Towards a Framework for Realizing Actionable Insight from Complex Data

A Machine-Augmented Cognition Approach to
Data Exploration, Information Synthesis, and Knowledge Actualization

SHIANG YEN, TAN

Submitted in fulfillment of the requirements for the degree of

Doctor of Philosophy

School of Electrical Engineering and Computer Science

Science and Engineering Faculty

Queensland University of Technology

2017

Abstract

Background: To solve complex analytics problems, data analysts often engage in a series of problem-solving activities, including extracting meaningful information from the data, synthesizing information to form higher-level concepts, creating a mental depiction of the problem, and imagining the impacts of possible scenarios. However, existing data analytics systems often focus exclusively on low-level data exploration and fail to effectively support these problem-solving activities. Consequently, analysts have to contend with the gaps between low-level technical analytic results and high-level conceptual understandings which are required to solve complex analytics problems. As a result, data analysts often find it is challenging to determine how the analytic results can be used to inform their decision-making and solve their real-world problems. In other words, existing data analytics systems often fail to deliver actionable insight.

Objectives: The goal of this study is to develop a design of data analytics systems that can explicitly support the analysts' problem-solving activities. This study theorized that when the problem-solving activities are supported, analysts are more likely to produce higher-level insights that can more readily inform decision-making. In order to achieve this design goal, this study asserts that there is a need for 1) systematically understanding and defining actionable insight, 2) understanding problem-solving activities and outcomes required to achieve actionable insight, and 3) proposing system features that can effectively support the problem-solving activities.

Method: This study employs design science research as the methodology to guide the overall design of this study. Through an integrated understanding of relevant theories, namely situation awareness (SA), sensemaking, and complex problem solving, this study conceptualizes actionable insight as a multi-component construct. Based on the way actionable insight is conceptualized, an explanatory framework is developed to provide a holistic explanation for complex analytics tasks. This framework is specifically contextualized in the field of data analytics to explain the information processes, user behaviours, cognitive states, and information artefacts in different phases of a complex analytics task. More importantly, this explanatory framework provides systematic and theoretically-grounded design requirements which can be leveraged to improve user performance in the problem-solving activities.

A design framework was then developed to provide a set of prescriptive design principles for how the design requirements can be addressed. The design framework also acts as the blueprint for translating the conceptual design into tangible system features. To evaluate the effectiveness of the proposed design, a user study involving 30 participants was undertaken in a controlled setting. A prototype system was developed based on the design framework. The prototype system was evaluated against a conventional data analytics system in the user study. The user study requires the participants to analyse stock markets and to develop stock portfolios that will maximize returns on investment.

Findings: This study categorizes the problem-solving activities into three phases: 1) data exploration, 2) information synthesis, and 3) knowledge actualization. The result shows that the proposed design is capable of enhancing the participants' performance in the information synthesis and knowledge actualization phases, but not in the data exploration phase. Additionally, the proposed design was found to increase the perceived quality of the analytical result, implying that the results are more likely to be deployed into the physical world through decision making. Lastly, mixed results were found on the effects of the proposed design on actual decision performance. Specifically, the qualitative aspect of the participants' decisions has been significantly improved, but the quantitative aspect of the decisions was not improved over conventional data analytics systems.

Overall, the findings suggest that the proposed design can support users to be more effective in integrating low-level technical findings into a holistic understanding of the problem situation and predicting and assessing the plausible impacts of the problem situation's future development. Such understandings that are meaningful at the problem-solving level reduce the gaps between the low-level technical analysis and high-level understanding required for solving analytical tasks. In comparison with conventional data analytics systems, the proposed design enables the users to derive analytics results that can more readily be used to inform decision making and to solve complex analytics problems. In other words, the proposed design improves the chance of deriving actionable insight from the data.

Conclusion: The contributions of this study include the explanatory framework which provides systematic understanding of analyst's workflow, behaviours, and information needs, as well as other design considerations, in three different phases of the data analytics process. The framework can be used by data analytics researchers to understand design considerations and requirements without repeatedly integrating the scattered knowledge from different domains. Additionally, the design framework can provide useful guidelines for practitioners to build data analytics systems that can effectively support the problem-solving activities of the users. As a further implication, it is hoped that the proposed data analytics system can help practitioners to harness greater value from their data, and thus can turn their IT investments into value-creation assets.

Keywords: Data analytics; Situation awareness; Sensemaking; Complex problem; Actionable insight, Information synthesis, Knowledge actualization; Stock investment analysis, Data Science

Acknowledgement

The completion of this thesis would not have become a reality without the invaluable support, encouragement, and inspiration of several individuals. Hence, I wish to present my appreciation to all those who have provided their supports in many different ways. Firstly, I would like to express my deepest gratitude to my principal supervisor, Dr. Taizan Chan, for his valuable guidance and support during this journey. I am deeply indebted as his constructive comments have helped to clear the cobweb and have kept me moving on the right track. I also wish to thank my associate supervisors, Dr. Ernest Foo and Associate Prof. Yue Xu, from the bottom of my heart for all their support, advice, and patience. I am very fortunate to study under their supervision.

I owe particular thanks to Jason Chung and Kang Jun, who have been willing to sacrifice their time in discussing and providing professional feedback on various aspects of stock investment analysis. Special thanks go to Nick Law and Dominic Paul for assisting in the participant recruitment process. I also want to thank Vincent Chia for his valuable advices on the programming matters, and more importantly, for being as a friend and comrade during this long journey. I also need to thank Jennifer Beale for providing thesis editing.

I would also like to express my gratitude to Queensland University of Technology (QUT), whose offer of the QUT Postgraduate Research Scholarship (QUT-PRS) has made it possible for me to complete this doctoral degree. Particular thanks also go to Elaine Reyes and Mrs. Prasanthi for their administrative support. The support they give to the HDR students is selfless and sincere.

Last but not least, I wish to thank my family members, especially my parents, for giving me emotional support and being my constant motivation to accomplish this thesis. I want to particularly thank my mom for her tireless support during the thesis writing phase. One person that I cannot thank enough is my fiancée, who has tirelessly shared all the burdens, the happiness, and the hardships with me.

Statement of Original Authorship

The work contained in this thesis has not been previously submitted to meet requirements for an award at this or any other higher education institution. To the best of my knowledge and belief, the thesis contains material previously published or written by another person except where due reference is made.

Signature: QUT Verified Signature

Date : 09 June 2017

Table of Contents

Abstract ..... **ii**
Acknowledgement ..... **iv**
Statement of Original Authorship ..... **v**
Table of Contents ..... **vi**
List of Tables ..... **xi**
List of Figures ..... **xiii**

Chapter 1 Introduction **1**
1.1 Background of Study: Data Analytics 1
1.2 Research Problems 4
1.3 Research Objectives and Central Approach 7
1.4 Outcomes and Significance 8
1.5 Scope of Study 9
1.6 Thesis Structure 11

Chapter 2 Literature Review **12**
2.1 Overview of Literature Review 12
2.2 Existing understanding of Actionable Insight 13
2.3 Situation Awareness (SA) Theory 16
2.3.1 Background of Situation Awareness (SA) Theory 16
2.3.2 What is Situation Awareness? 17
2.3.3 Situation Awareness Level-1 (SA1) 18
2.3.4 Situation Awareness Level-2 (SA2) 19
2.3.5 Situation Awareness Level-3 (SA3) 21
2.3.6 Cognitive Obstacles to achieve SA 23
2.3.7 Situation Awareness, Decision Making, and Performance 24
2.3.8 Mapping between Situation Awareness and Insights 25
2.4 Sensemaking Theory 26
2.4.1 Background of Sensemaking 27
2.4.2 Sensemaking and Complex Problem Situation 27
2.4.3 Understanding Sensemaking 28
2.4.4 Mapping between Sensemaking Theory and Situation Awareness 32

2.5 Complex Problem Situation 34
2.5.1 Characteristics of the Complex Problem Situation 34
2.5.2 Complexity 35
2.5.3 Connectivity 36
2.5.4 Uncertainty 37
2.5.5 Dynamicity 38
2.5.6 Summarized Key Points from Complex Problem Situation 39
2.6 Summary: A Big Picture of the Justificatory Theories 40
2.7 Related Works 42
2.7.1 Commercial Products 42
2.7.2 Research Works 44
2.7.3 Summary of Review on Related Works 47
2.8 Summary of Literature Review 49

Chapter 3 Research Methodology & Design 51
3.1 Overview 51
3.2 Design Science Research as the Central Research Methodology 51
3.3 Research Design 53
3.4 Research Activities and Flow 54
3.5 Data Collection & Analysis Methods 61

Chapter 4 Developing the Conceptual Explanatory Framework 62
4.1 Overview 62
4.2 Conceptualizing Actionable Insight and its Components 63
4.3 Defining Actionable Insight 66
4.4 The Hierarchical Framework of Insights 67
4.4.1 Major Component 1: Analytic insight 68
4.4.2 Major Component 2: Synergic Insight 74
4.4.3 Major Component 3: Prognostic Insight 82
4.4.4 Summary of the Hierarchical Framework of Insight 93
4.5 Summary of Design Requirements 95

Chapter 5 Developing the Conceptual Design Framework 97
5.1 Overview 97
5.2 Design Philosophy 98
5.3 Design Principles and Operationalization 100
5.3.1 Enabling Divergent Exploration 101
5.3.2 Enabling Managed Observations 109
5.3.3 Enabling Exploration Convergence 119
5.3.4 Enabling Knowledge Creation 124
5.3.5 Enabling Assisted Situation Modeling 131
5.3.6 Enabling Predictive Reasoning 142
5.3.7 Enabling Stochastic Optimization 148
5.4 Conceptual Design Framework 155

Chapter 6 Designing the Evaluation 161
6.1 Overview of Evaluation 161
6.2 Requirements for the Evaluation 161
6.2.1 Needs for measuring the actual performance 161
6.2.2 Needs for controlling extraneous variables 163
6.2.3 Needs for assessing the design as an information system 164
6.2.4 Summary of the Evaluation Requirements 165
6.3 Operationalization of Constructs 165
6.3.1 Construct 1: Extents of Insight 165
6.3.2 Construct 2: Value of Analytic Outcome 169
6.3.3 Construct 3: Decision Performance 172
6.3.4 Construct 4: Usability 174
6.3.5 Construct 5: Cognitive Load 176
6.4 Designing the User Study 178
6.4.1 Participants 178
6.4.2 Activities in the User Study 179
6.4.3 Task Design 180
6.4.4 Data Analysis Method 182
6.4.5 Pre-test 182

Chapter 7 Evaluating the Designs 184
7.1 Overview 184
7.2 Demographics of the Participants 184
7.3 Evaluating the Main Hypotheses 185
7.3.1 Evaluating the Effects on the Insight Components 185
7.3.2 Evaluating the Effects on the Value of Analysis Outcomes 188
7.3.3 Evaluating the Effects on Decision Performance 191
7.3.4 Summary of Hypotheses Testing 193
7.4 Evaluating the Designs as an Information System 194
7.4.1 Usability 194
7.4.2 Cognitive Load 195
7.5 Discussion on the Analysis Results 196
7.5.1 Findings pertaining to the level of insights 196
7.5.2 Findings pertaining to the value of the analysis outcome 203
7.5.3 Findings pertaining to the decision performance 208
7.5.4 Findings pertaining to Usability 208
7.5.5 Finding pertaining to Cognitive Load 211
7.5.6 Relationships from Insights to Decision Performance 213
7.6 Implications of the results on the Propositions 214

Chapter 8 Conclusion 216
8.1 Overview 216
8.2 Research Investigation 216
8.2.1 Answering Research Question: How can actionable insight be systematically defined? 218
8.2.2 Answering Research Question: What are the processes and requirements to achieve actionable insight? 219
8.2.3 Answering Research Question: How can the processes and requirements be effectively supported? 220
8.3 Contributions 223
8.3.1 A Definition of Actionable Insight 223
8.3.2 Understanding of Complex Data Analytics Tasks 224
8.3.3 Design Recommendations for Data Analytics Systems 225
8.3.4 An Implementation Reference 225
8.4 Limitations and Future Works 226
8.5 Final Remarks 228
Reference 229
Appendices 240

10.1 Appendix A – Questionnaire Items 240
10.2 Appendix B – Ethic Clearance Approval 246

List of Tables

Table 1. Comparing two types of data analytics systems 2
Table 2. Major causes of failure to obtain perceptive awareness 19
Table 3. Characteristics of complex problem situation 35
Table 4. Components of a design theory 51
Table 5. Summary of research design 54
Table 6. Details of research activities 56
Table 7. Argument types 77
Table 8. Variations in hypothesized scenarios 85
Table 9. Possible elements of a variation 89
Table 10. Summary of design requirements 95
Table 11. Components of a design principle 100
Table 12. The information which can be used for meta-observation analysis 123
Table 13. Information can be extracted from the texts 129
Table 14. A simple example of uncertainty in a situation model's factors 149
Table 15. Summary of the design principles 158
Table 16. Analytical performance at three different stages 163
Table 17. Propositions derived from the conceptual design framework 166
Table 18. An overview of the measurement for extents of insights 167
Table 19. Variables for levels of insight 168
Table 20. An overview of measurement for Value of Analytic Outcome 169
Table 21. Dimensions for perceived value of analysis outcomes 170
Table 22. Summary of the measurement for decision performance 172
Table 23. Variables and their measurement for decision performance 173
Table 24. An overview of measurement for usability 174
Table 25. Variables and measurements for usability 174
Table 26 Variables of Cognitive Load 177
Table 27. Demographics of the participants 184
Table 28. Paired-sample t-test on the overall insight 186
Table 29. Paired T-tests on the six dimensions of insight 187
Table 30. Paired-sample T-test on value of analytic outcome 188
Table 31. Paired-sample T-test on dimensions of the value of analytic outcome 189
Table 32. Paired-sample T-test on total earnings 191
Table 33. Paired-sample T-test of the earnings above the random baseline 192
Table 34. Summary of the main hypothesis testing 193
Table 35. Paired sample T-tests on dimensions of usability 194

Table 36. Paired-sample T-test on the dimensions of cognitive load 195
Table 37. Components of actionable insight..... 218
Table 38. Components of actionable insight and their problem-solving activities..... 220
Table 39. Design principles and problem-solving activities..... 221

---- This space is intentionally left blank ----

List of Figures

Figure 1. Consequences of the lack of supports for problem-solving activities 5
Figure 2. Causes for existing data analytics in failing to delivery actionable insight 6
Figure 3. Data Analytics as a broad concept and the focus of this study 10
Figure 4. Type of problem targeted by this study 10
Figure 5. Types of insight in surveyed literature 14
Figure 6. Types of insight and abstraction level 16
Figure 7. Situation awareness, decision, and performance 24
Figure 8. Commonalities between situation awareness and insights 26
Figure 9. Sensemaking model in business analytics 29
Figure 10. Connections between the sensemaking theory with SA theory 33
Figure 11. Overview of justificatory knowledge in this study 41
Figure 12. Three views in ARUVI: data view, navigation view, and knowledge view 45
Figure 13. SRS's interface for information synthesis 46
Figure 14. Analysis of hypothesis network in SRS 47
Figure 15. Phases in data analytics and supports 49
Figure 16 Flow of research activities in design research 53
Figure 17 Research activities and flow 55
Figure 18 Contents of Chapter 4 and Research Objectives 63
Figure 19. Major types of insight and the states of situation awareness 64
Figure 20. Insights in Data Analytics 65
Figure 21. Simplified representation of the conceptual explanatory framework 68
Figure 22. From dataset to analytic insight: identification + perceptive insights 69
Figure 23. Identification insight involves search & filter activity 70
Figure 24. Perceptive insight involves perceive & interpret activity 71
Figure 25. Conceptual structure of an observation 72
Figure 26. Characteristics of observations in complex problem situation 73
Figure 27. From analytic insights to synergic insight: integrative + comprehensive insights 75
Figure 28. Integrative insight involves integration and synthesis 76
Figure 29. Conceptual structure of an argument / association 78
Figure 30. Comprehensive insight involving connect & build activity 79
Figure 31. Prognostic insight: predictive + prescriptive insights 83
Figure 32. Predictive insight involves predict & simulate 85
Figure 33. Prescriptive insight involving optimization and prescription 89
Figure 34. Summary of HIVE framework 93
Figure 35. Characteristics of the three major insights 94

Figure 36 Contents of Chapter and Research Objective 97
Figure 37. Relative position of "machine-augmented cognition" approach 99
Figure 38. Dimensions of data divergence: aspects and levels 102
Figure 39. Supporting data divergence with dataset integration 103
Figure 40. Support divergence enquiry with interactive multi-modal enquiries 104
Figure 41. Interfaces for pulling data from multiple repositories 104
Figure 42. Enable users to integrate and link data with no technical skills required 105
Figure 43. Multi-modal enquiry enabled by visual analytics 106
Figure 44. Enquiries are viewed next to each other 107
Figure 45. Supporting the observations to be systematically stored, managed, and retrieved 111
Figure 46. Capturing the state of an enquiry 112
Figure 47. Interface for capturing an observation 113
Figure 48. User-defined key attributes of an observation 114
Figure 49. Interface for managing the attributes 114
Figure 50. Examples of attribute type can be defined by users 116
Figure 51. List of observations made 117
Figure 52. Search and filter functions in the observation list 118
Figure 53. Converging observations 120
Figure 54. Converged view in overall 121
Figure 55. Wizard for customizing the visualization in the converged view 122
Figure 56. Enabling knowledge creation by integrating observations and synthesizing reasoning 126
Figure 57. Selecting observations to be associated with an argument 127
Figure 58. Enabling reasoning to be captured as the attributes of argument 129
Figure 59. Text analysis extracting user's reasoning into structural information 130
Figure 60. Supporting the users to identify the core structure of a situation model 134
Figure 61. Supporting the situation modeling with both quantitative and qualitative approaches 136
Figure 62. Enabling users to start the situation modeling with some established templates 138
Figure 63. Interface for situation modeling 139
Figure 64. Dynamic situation model 140
Figure 65. Relationships between the factors 140
Figure 66. Supporting predictive reasoning 145
Figure 67. Enabling prediction or posing what-if question 146
Figure 68. Prediction facilitated by forecasting algorithms 147
Figure 69. Enabling optimization and risk assessment 151
Figure 70. Interface for optimization and risk assessment 152
Figure 71. Enlarged view of the key information in box 1 152
Figure 72. Slider and input field for interaction 153

Figure 73. Pop-up dialog for tweaking the optimization and risk assessment 153
Figure 74. Enabling users to assess the risk of their own resource allocation plan 154
Figure 75. Conceptual design framework 157
Figure 76. Performance at three different stages 162
Figure 77. Processes in data collection 164
Figure 78. Activities in each session of user study 179
Figure 79. Data types included in the user study's datasets 181
Figure 80. Randomization of the session sequence and datasets 182
Figure 81. Insight, its dimensions, and the tests 186
Figure 82. Proportion of participants with positive and negative returns 191
Figure 83. Proportion of participants above and below the random average 192
Figure 84. Match percentage between the treatment and control group 193
Figure 85. Conceptual position of insights 196
Figure 86. Propositions and Results 214
Figure 87. Integrative framework for technical data analysis techniques 226

Chapter 1
Introduction

1.1 Background of Study: Data Analytics

Humans have created a digital universe. Like the physical universe, the digital universe of data is enormous and is doubling in size every two years. By 2020, this man-made universe will contain nearly as many digital bits as there are stars in the universe. In number, the data humans create and share annually will reach 44 trillion gigabytes in less than 1,500 days (IDC, 2014). Today, a modern institution handles more than 60 terabytes of data annually, which is about 1,000 times more than a decade ago (Beath, Irma, Ross, & Short, 2012). The exponential growth of the data can be attributed to the proliferation of the Internet of Things (IoT), mobile devices, social media, and personalized web experience. The highly complex, heterogeneous, and dynamic data requires new breeds of analytics tools to unleash the opportunities in the enormous volume of data.

Data Analytics (DA), a discipline that arises to meet the needs for greater analytic capabilities, involves the use of visualizations, statistics, mathematical models, and machine learning to extract useful knowledge from large and complex data. The ideology of data analytics is to enable data-driven insights that can inform decision and device action (Stubbs, 2011), generally known as *actionable insight*. In the industry, the term *actionable insight* has been widely used by business executives, consultants, and software vendors as the goal of Data Analytics. More importantly, actionable insight is the key driver for businesses to invest in Data Analytics solutions (Sawyer, 2011).

As a broad discipline, Data Analytics consists of three major areas, data collection, data storage, and data analysis. The focus of this study is the "data analysis" area, the data analytics, in which the human-information discourse occurs to derive *actionable insights* from the data. This is because what is crucial in the end is not how much data can be collected and stored; it is more about what users can do with the data that counts. Therefore, the real value of Data Analytics is hinged on the ability to analyze the data. Moreover, an analysis gap exists because the capability to collect and store data has been growing rapidly, but the capability to analyze these data increases at a much slower pace (Keim, Mansmann, Schneidewind, & Ziegler, 2006). This study's focus on the data analysis area will allow this study contributes as a part of the endeavor to close the analysis gap.

Data analytics has been seen as a necessity for modern institutions to leverage the data universe for delivering real value. Institutions capitalize on data analytics to improve decision making in diverse domains, including business, finance, healthcare, emergency services, environmental policy, and scientific research. Data analytics provides users with the unprecedented capability to deal with massive

and complex data. While conventional data analysis has been focusing on answering retrospective questions such as what, where, when, and how many, data analytics goes beyond that to gain deeper understandings of why the problem happening is, to make predictions, and to optimize actions to be taken. Then, institutions can proactively allocate resources and formulate action plans to maximize their goal with greater resource efficiency. Ideologically, data analytics enables institutions to translate their massive data repositories into actionable insight that delivers pragmatic value.

The data analytics process often involves using one or more data analytics systems. Data analytics systems can be categorized into two types based on their approaches: 1) visual analytics, which capitalizes on interactive data visualization for ad-hoc querying and data discovery and 2) computational analytics that relies on statistical and computational techniques such as predictive analysis, data mining, and artificial intelligence to extract key information from large datasets. These two system types serve different purposes, have different processing capacities, and require different levels of skill to operate. *Table 1* provides a comparison view of the two types of data analytics systems.

Table 1. Comparing two types of data analytics systems

	Visual Analytics	Computational Analytics
Purpose	▪ To gain deeper, qualitative insight from the data	▪ To extract hidden knowledge from massive data
Input Data Structure	▪ Structured to semi-structured data	▪ Structured to non-structured data
Data Processing Capability	▪ Low to Medium	▪ High to Enormous
Main Approach	▪ Human-driven sensemaking and reasoning	▪ Machine-driven computations and modeling
Techniques	▪ Interactive visualization ▪ Visual and structural thinking	▪ Statistical and mathematical modeling ▪ Artificial intelligence and machine learning
User Friendliness	▪ High (Easy to use)	▪ Low (Difficult to use)
Example of Systems	▪ SAS Visual Analytics ▪ SAP Visual Intelligence ▪ IBM Cognos Analytics ▪ Tableau (Stanford University)	▪ SAS Enterprise Miner ▪ RapidMiner Studio ▪ IBM SPSS Modeler and SPSS Statistics ▪ R statistical library

(IBM Global Business Service, 2010; Liebowitz, 2013; Surma, 2011)

Visual analytics systems capitalize on interactive visualization techniques to facilitate human analysts in understanding data, in reasoning, and in decision making (Keim, 2012). Visual analytics strives to empower non-technical users with the analytical capability via user-friendly interfaces (Eckerson, 2009). The ideas of self-service and pervasive analytics enabled by visual analytics systems extend its user pool beyond the well-trained data scientists and analysts to general users who

traditionally rely on others for data analytics. Self-service analytics enables the ideal situation where the person who needs the insight is the same person who analyzes the data. This is desirable because the person who possesses the domain knowledge can better harness the true value of the data (Mirel, 2004). As a result, visual analytics gives greater flexibility to the users to redefine their information seeking strategy on the fly, allows them to react more quickly to their problem, and eventually shortens the time from analysis to value.

On the other hand, computational analytics systems are a group of machine-driven analysis tools which relies on mathematical, statistical, data mining, and machine learning techniques to automate knowledge discovery (Surma, 2011). Computational analytics can handle enormous sets of complex data and can execute the analysis in an exhaustive and systematic manner (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Such capability is especially important for today's data, which exceeds the human's attention and cognition capacities. Additionally, computational analytics can take advantage of the advancement in computer technologies, such as neurocognitive chips, in-memory database, and quantum computing, to achieve unprecedented data processing capability.

The powerful computational analytics is not without its downside. The knowledge discovered by computational analytics is mostly 'passive' knowledge. Such passive knowledge is often highly technical and with little domain context (Cao, 2012). Such knowledge is also often too uncertain to interpret from the observable part of the real world (Ohsawa & Nishihara, 2012). As a result, the practical value of such knowledge is usually trivial, as it tells the decision makers very little about how to act upon it. One of the causes is that the computational analytics often involves pure computations in a black-box setting and does not factor relevant domain knowledge into the knowledge discovery process (Keim, Kohlhammer, Ellis, & Mansmann, 2010). Moreover, it is widely agreed that computational analytics systems are too complex for widespread use because highly specialized knowledge and skills are required to operate them (Nemati, Earle, Arekapudi, & Mamani, 2010). The statistic shows that only less than 17% of institutions have been actively using computational analytics to support their operations.

In contrast, whilst *visual analytics* systems are relatively easier to use, their processing capability is significantly limited by the human user's capacity. This is because visual analytics relies on manual efforts to derive insights through manipulating data representations, such as graph, chart, scatter plot, map, gauges, and dashboard. Given its reliance on the human analysts to perceive patterns and trends in the visualizations, visual analytics is generally more time consuming, more susceptible to human bias, and the process cannot be automatically improved through steadily growing computational resources (Keim, Mansmann, Schneidewind, Thomas, & Ziegler, 2008). Nevertheless, visual analytics allows users' domain knowledge to be incorporated to develop domain-meaningful visualizations and to drive the information seeking efforts. This characteristic of visual analytics is especially important

for complex and strategic problems in which human discretion and judgmental heuristic are critical for deriving the solutions. In short, an effective solution to the analytic problem in the real world requires a good balance between the powerful computational analytics and the context-aware visual analytics.

1.2 Research Problems

The purpose of data analytics is twofold. Firstly, data analytics supports users to understand their data. Secondly, its higher-order purpose is to solve the users' real-world problems through informed decision making. For example, data analysis on social media contents, previous marketing strategies, and the company's reputation first enables the users to understand the states of these factors and the relationships between them, then the situational understandings enable the users to devise the optimal combination of marketing strategies for improving the company's image in a public relations crisis. In this notion, the true goal of data analytics is to support users to make well-informed decisions that solve pragmatic problems. In other words, the true objective of users is to gain actionable insight from the data analytics process. By its coarse definition, actionable insight refers to the information which has the sufficient depth and breadth to be the basis for the users to take on action to solve their real-world problems, which in turn yield positive returns.

The real-world analytics problems are often complex and ill-structured. By complex, it implies that the problems often contain a huge number of interconnected data elements which require the powerful processing capability of computational analytics. On the other hand, its ill-structured nature implies the criticality of domain knowledge and heuristic judgment from the human users to determine the context, meaning, and relations in the analysis. It is well-recognized that the complex analytics problems generally require users to engage in a series of problem-solving activities (David & Michelle, 2009; J. Kohlhammer, Keim, Pohl, Santucci, & Andrienko, 2011). These activities include extracting information from data, synthesizing the information into a big picture, creating a conceptual depiction of the problem, imagining the impacts of different possible scenarios, and generating potential solutions. These problem-solving activities enable users to gain different levels of understanding of their problem. These understandings provide the users with the levels of depth and breadth about the problem which are required by the users to achieve actionable insight. In other words, the problem-solving activities are the enablers of actionable insight. Prior study has also shown that higher levels of situation understanding are the key to effective problem solving and are often found to be positively associated with higher decision quality (Yadav & Khazanchi, 1992).

Nevertheless, many existing data analytics systems are not designed for effectively supporting these problem-solving activities (Mirel, 2004). Substantial data analytics systems to date are the results of advancement in data-driven and computational-oriented techniques. They were designed with little consideration of how users behave in a complex data analytics task and how they can be supported to

perform better. More importantly, these systems often support users only in the low-level problem-solving activities such extracting information from their data, but not in the high-level complex problem-solving activities such as synthesizing the extracted information to form concepts that are significant at the problem level, creating an integrated view of the problem from the concepts, and assessing the impact of various scenarios and potential actions. Without these high-level activities being supported, users are lacking of the high-level information with sufficient depth and breadth required to devise solution for their problems. Therefore, there exists a gap between the low-level technical analytic results and the high-level conceptual understandings required to solve the complex analytics problem. For an instance, low-level analytics results, such as sales prediction, demand forecasting, and social media sentimental analysis, do not lead to a solution for the social media crisis that a company faces. These low-level information needs to go through a series of high-level problem-solving activities to join the missing links between them, to contextualize them in the specific scenario based on the user domain knowledge and judgment, and to produce conceptual understanding of the problem that allows the users to devise potential solutions.

As the consequence, practitioners often find it is challenging to determine how the data analytic results can be used to inform their decision-making and solve their real-world problem (Harris, 2005; Houxing, 2010; Stijn & Annabel Van den, 2011). Researchers have also been commonly critical because the existing data analytics systems that focus on low-level tasks do not map well to the true **goal of a**nalysts (Amar & Stasko, 2005; Tim, 2006). Indeed, it has been commonly reported that results from the data analytics systems often have little value to domain experts (Cao, Luo, & Zhang, 2007; Saraiya, North, & Duca, 2004). In other words, these data analytics systems have failed to achieve the higher-order purpose of solving problems through informed decision making. *Figure 1* summarizes this paragraph by illustrating the consequences of the lack of support for problem-solving activities.

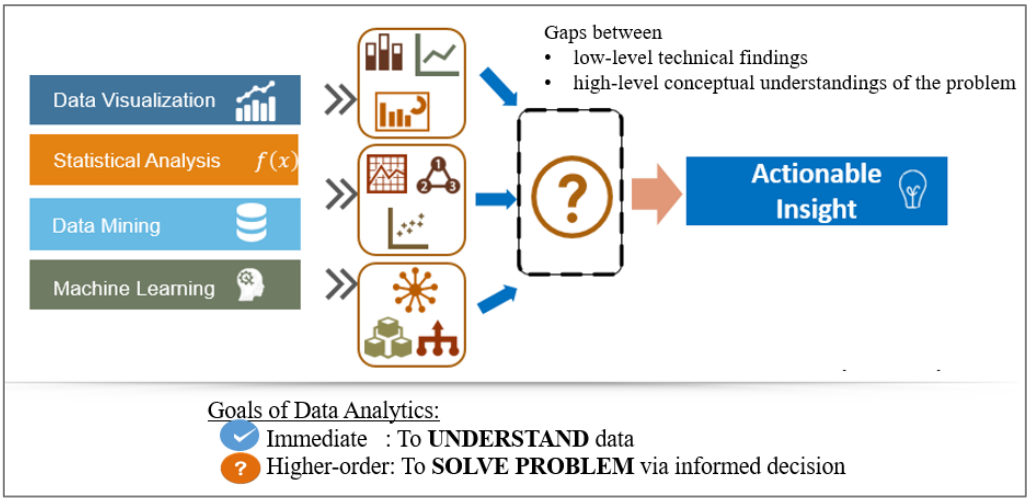


Figure 1. Consequences of the lack of supports for problem-solving activities

In the notion of data analytics' ideology, these systems have failed to deliver *actionable insight*. This study asserts that the crux of the problem lies in the fact that existing systems do not effectively support users to carry out the problem-solving activities required to solve complex analytics problems. This study suggests that the cause of this ineffective design is the lack of clear understanding of what are the problem-solving activities that require support. Most of the research on data analytics and their resultant frameworks have been emphasized on the technical and computational aspects, and thus have significantly improved the low-level problem-solving activities, such as perceiving and interpreting data from visualizations (Green, Ribarsky, & Fisher, 2009; Heer & Shneiderman, 2012; Jankun-Kelly, Kwan-Liu, & Gertz, 2007). In contrast, there is lack of holistic theories or frameworks that focus on the entire data analytics process, from perceiving data, synthesizing information, assessing the hypotheses, to the realization of analytics solutions. There is a need for a comprehensive understanding of the process and requirements required for the users to achieve actionable insight. A framework or theory that is built upon established theoretical foundations, such as cognition and sensemaking, is needed for the data analytics process (Endsley, Bolte, & Jones, 2011).

Such a lack of understanding of the process and requirements is in turn caused by the lack of a systematic and theory-driven understanding of what actionable insight is. Despite the term's popularity, there is a lack of common agreement on what exactly is actionable insight. Most if not all publications or reports found have used the term "actionable insight" without formally defining it (Basole, Hu, Patel, & Stasko, 2012, Burby & Atchison, 2007). The highly abstract and vague understanding of actionable insight impedes the progress of works on the design and evaluation of effective data analytics systems. Without clearly knowing what the desired outcome is, it is difficult to know how the users can be supported effectively and to know whether the systems live up to their claims. There is a need for a definition of actionable insight that is systematic, theoretically grounded, and measurable. This chain of causes and effects is illustrated in *Figure 2*.

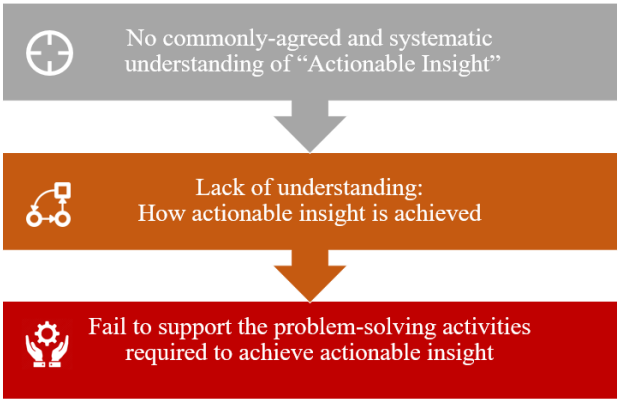


Figure 2. Causes for existing data analytics in failing to delivery actionable insight

The is a need for the understanding of actionable insight that can be used to 1) systematically define actionable insight, 2) comprehensively understand the process and requirements to achieve actionable insight, and 3) inform the design of data analytics systems that can effectively support the user's complex problem-solving activities. Therefore, these pose the three main research questions that drive this study. The following are three research questions of this study.

1) ***How can actionable insight be systematically defined?***

2) ***What are the processes and requirements to achieve actionable insight?***

3) ***How can these processes and requirements can be effectively supported?***

1.3 Research Objectives and Central Approach

To answer the research questions, the following research objectives are formulated:

- ***Objective A***: To propose a systematic and theory-driven definition of actionable insight
- ***Objective B***: To develop a conceptual framework for understanding the processes and requirements of actionable insight
- ***Objective C***: To create a design for a data analytics system that supports the processes and requirements
- ***Objective D***: To evaluate the proposed design of the data analytics system

This study employs *design science research* as the methodology to guide the overall design of this study. A literature review was first undertaken to assess existing works on data analytics systems for complex analytics problems. The literature review also seeks to understand relevant theories and concepts that can explain the processes and user behaviors in a data analytics task.

Based on the integrated understanding gained from the relevant theories, this study conceptualizes actionable insight as a multi-component construct. Based on the way this study conceptualizes actionable insight, a systematic and theoretical-driven definition of actionable insight is proposed. The definition addresses ***Research Objective A***.

Subsequently, a conceptual explanatory framework was developed to provide a holistic explanation for the complex analytics task. This framework includes the processes, user behaviors, information artefacts, and cognitive outcomes in different phases of data analytics process. More importantly, the

conceptual explanatory framework specifies a series of design requirements that need to be fulfilled in order to enable data analytics systems to effectively support the complex problem solving activities of users. Thereby, the conceptual explanatory framework addresses ***Research Objective B***.

Based on the design requirements, a conceptual design framework was developed to give an integrated view of how the design effects work together. More importantly, the framework contains a set of design principles in which each of them provides prescriptive design statements on how a corresponding design effect can be achieved. The design principles act as the detailed blueprint for translating the conceptual design into tangible system features. This conceptual design framework, therefore, addresses ***Research Objective C***.

To evaluate the effectiveness of the proposed design, a user study was undertaken in a controlled setting. A prototype system was developed based on the design principles. The prototype system was evaluated against a conventional data analytics system. Subsequently, the results from the user study were used to reflect on the design principles and the overall design framework. The evaluation, therefore, addresses ***Research Objective D***. At the end of this study, the primary outcome is the validated design knowledge for building data analytics systems which can effectively support the users' problem-solving activities along the data analytics process.

1.4 Outcomes and Significance

Four main research outcomes are produced at the end of this study: 1) a definition of actionable insight, 2) an explanatory framework for understanding the complex data analytics tasks, 3) a design framework for designing a data analytics system that can effectively support users to solve complex analytics problem, and 4) the prototype system as an instantiation of the design. These outcomes are important as to:

1) **Establish a shared and systematic understanding of actionable insight**

   Being able to systematically understand actionable insight is important for 1) evaluating whether a data analytics system has lived up to its claims in term of what it supposes to deliver and 2) to inform the developers and researchers about the qualities of insights that are seek by the users, thus allowing them to design the systems that can effectively support these qualities. Therefore, understanding of actionable insight could be the first step towards more effective data analytics systems.

2) **Understand the complex analytic task in a holistic manner**

   The explanatory framework describes the user behaviors, cognitive states, interaction outcomes, and other considerations in different phases of a complex data analytic task. Such understanding allows the leverage points where the user analytical performance can be enhanced to be

identified as the design requirements. The framework can be used to informed the design, or to be improved upon by future researchers in data analytics area without the need for reinventing the wheel.

3) **Recommend a validated design of data analytics system which can effectively support the users in the data analytics process.**

   While most existing data analytics systems focus technical aspects such as data manipulation and data transformation, this study recommend a holistic data analytics framework that goes beyond support low-level data manipulation to support the high-level problem-solving activities such as integrating technical findings into knowledge, creating a big picture of the problem, and imagining the impacts of possible scenarios. This study asserts that by supporting these problem-solving activities, the gap between the low-level findings and the high-level understandings required to solve the analytics problem can be reduced, thus allowing the users to better achieve actionable insight.

4) **Provide an implementation reference for the conceptual design**

   The prototype system developed in this study is a detailed reference of how the abstract design can be translated into tangible systems. It helps the audiences to have a solid understanding and to better appreciate the theoretical aspect of data analytic framework. Moreover, it is useful as a tangible example that can be referred by practitioners such as software engineers to derive specific system functionalities from the conceptual architecture being proposed. In short, the prototype system increases the pragmatic value of this information system study by reducing the gaps between the theoretical articulations and the readiness to being deployed in the real world.

As a further implication, it is hoped that the proposed design of data analytics system can help practitioners to increase the insight throughput from the data analysis process, consequently allowing them to harness greater value from their enormous data, and thus, turning the IT infrastructure and investments such as data warehouse, data mining, online analytical processing (OLAP) into value-creation assets.

1.5 Scope of Study

In order for a study to be meaningful, it is critical for the scope of the study to be specific, manageable, and realistic. This information system study is positioned as a study in the area of data analytics. As a broad concept, Data Analytics consists of three major phases: data collection, data storage, and data analysis. This study focuses on the data analysis phase, which is the most critical phase which transforms data into actionable insights, and yet its advancement is relatively slower compared

to the two other phases. This focus also leads to an assumption that the data to be analyzed has already gone through proper collection, integration, management, and storage processes. For instance, this study assumes that all the data required for the analysis is able to be acquired and to be readily accessed from the databases. Figure 3 shows the scope of this study.

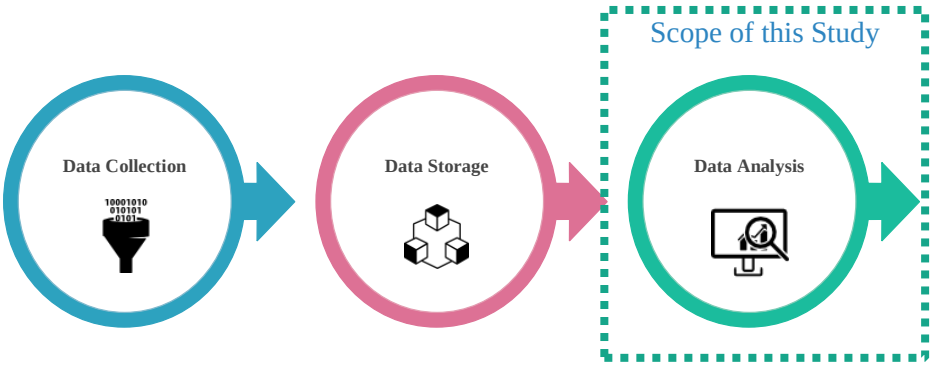


Figure 3. Data Analytics as a broad concept and the focus of this study

Different types of analytics problems require different data analytics systems to handle them. The analytics problems can lie anywhere on a continuum with extremely ill-structured problems at one end and with extremely well-structured problems at another end. Well-structured analytics problems are often routine problems of which the processes to solve the problem is clear and can be known beforehand. These problems can be automated with minimal or no human intervention. In contrast, ill-structured problems are often strategic- or tactical-level problems in which the process to solve the problem cannot be predetermined and must be discovered in the process of analysis. This study focuses on the ill-structured problems, which may not be not entirely new or unique, yet every case has its considerable degree of uniqueness. The problems have to be solved on a case-by-case basis and inevitably require the domain knowledge, experience, and heuristic judgment of the human users. Examples of these problems include crime investigation, counterterrorism analysis, disaster damage control, public policy making, and financial investment decision. These are the types of analytics problems commonly faced by practitioners.

Figure 4. Type of problem targeted by this study

1.6 Thesis Structure

Chapter 2 presents the literature review which was undertaken to understand existing works on the systems for complex analytics problems. The literature review also includes relevant theories and concepts that can be used to explain the processes and user behaviors in a complex analytics task.

Chapter 3 outlines the research methodology, the research design, and the processes of this study. The research design describes the study's characteristics in terms of purpose, research setting, time, unit of analysis, and type of analysis. The processes specify the flow of research activities used to achieve the research objectives in this study.

Chapter 4 describes the conceptual explanatory framework which was developed to provide a holistic explanation on complex analytics tasks. This framework explains the processes, user behaviors, and cognitive states, and reasoning outcomes in different phases of data analytics. The framework provides design requirements for improving the design of the data analytics systems.

Chapter 5 presents the conceptual design framework developed based on the design requirements. The conceptual design framework contains a set of design principles in which each of the design principles describes how a corresponding design requirement can be achieved.

Chapter 6 describes how the design is being evaluated. The evaluation involves a user study which was undertaken in a controlled setting. This chapter illustrates the participant recruitment process, the task designed to be carried out by the participants, and the way participants are assigned to different user groups.

Chapter 7 covers the results and discussions from the evaluation. The main focus of this chapter is on the results and discussions whether or not the proposed design can effectively enhance the participant's analytical performance. The chapter also includes the evaluation on the usability, learnability, and effort required of the proposed system.

Chapter 8 highlights the study's findings and contributions and concludes the study by discussing the limitation and potential future works stemming from this study.

Chapter 2
Literature Review

2.1 Overview of Literature Review

The motivation that drives this study is the limited ability of the existing data analytics systems to help users to gain "actionable insight". Actionable insight is generally known as the level of understanding of the analytics problem that is sufficient for users to make informed decisions for solving the problem in practice. As noted in the *Research Problems* section (Section 1.2), problems in practice are mostly complex and ill-structured. These complex analytics problems require users to engage in a series of problem-solving activities. However, most of the existing data analytics systems fail to support these problem-solving activities required for users to achieve actionable insights. In other words, although existing data analytics systems can effectively support users' immediate goals to explore large and complex datasets, these systems often fall short in supporting users' higher-order goal to solve the problem.

To ground the understanding of this research problem in empirical studies, the existing definitions of actionable insight are first reviewed in Section 2.2. The review unveils that actionable insight can be understood as the product of different awareness states which the users gain along a problem-solving process. With the understanding that different awareness states need to precede the occurrence of actionable insight, situation awareness (SA) theory is reviewed in Section 2.3 to understand what the awareness states users can gain in a problem-solving process.

While SA theory provides an *outcome perspective* of a problem solving, Section 2.4 reviews the sensemaking theory that provides a *process perspective* of the problem solving. Sensemaking theory describes the information processing activities required to achieve the awareness states. This perspective provides useful insights into how users process the information, and thus can potentially reveal the leverage points in the process where user performance can be enhanced. In Section 2.5, this study seeks to systematically understand the *environment perspective* of the complex problem-solving task.

Section 2.6 summarizes the reviewed theories and concepts into a holistic frame. Then, in Section 2.7, relevant academic and commercial works to ensure the originality of this study's work. The section also describes the differences between the proposed data analytics systems in this study and existing works. Lastly, Section 2.8 summarizes the findings of the literature review.

2.2 Existing understanding of Actionable Insight

The term "actionable insight" has gained much attention in both industry and academia in the past 5 years. In industry, business executives, consultants, and software vendors have widely described actionable insight as the deliverable of data analytics software. System vendors often use the term in their slogans, marketing media, and white papers. In academia, the term has been used in publications from different disciplines such as data mining, information visualization, business analytics, and psychology.

Despite the popularity of the term, it has often been used without definition. For example, an IEEE publication stated that "the idea of a crystal ball that provides capabilities to explore, make sense of, and perhaps even provide actionable insight into rapidly changing…" (Basole, Hu, Patel, & Stasko, 2012). An example from a different source stated that "the real issue is that companies aren't using the tools to gather actionable insight, and they're not prepared to act on those insights…" (Burby & Atchison, 2007). No clear definition or explanation of actionable insight is given in the rest of the publications reviewed.

There is no systematic or formal definition of actionable insight. The understanding of actionable insight is varied across the different publications or sources. More importantly, the definitions are often too vague and too abstract for stimulating solid academic conversations, guiding the design of data analytics systems, or evaluating the data analytics performance. The ambiguous and highly abstract understanding of the term impedes the development of relevant measurement instruments. To the best of this study's effort, no measurement model or instrument for actionable insight or similar concept has been found. Added to this situation is that most of the existing works have been focused on measuring immediate outcomes of data analytics, such as number of patterns discovered, number of association rules extracted, and number of new knowledge learnt from visualizations. These performance measurements have dominated the system evaluation as its countable and objective nature is desired in computer science works which often focus on a specific technique. Although these measurements are good for evaluating the technical effectiveness of a specific technique, they often fail to measure the real goal of data analytics: how effective the data analytics systems can help users to gain meaningful and relevant information that allow them to solve their analytics problems. The importance of measuring the system performance at the level of user's goal has only gained popularity in the recent years (Cao, 2012). Actionable insight is regarded as a potential measure for the high-level goal.

By reviewing articles in analytic-relevant domains, this study found that the two common elements of actionable insight: 1) actionable insight is the outcome gained by analyzing information that can be used to make informed decision or to take action (Schneider & Gibson, 2011) and that 2) actionable insight allows users to achieve a positive value or a desired goal in their domain (Bose, 2009; Tim,

2006). Based these common elements, this study conjectures that actionable insight is the most comprehensive understanding of the analytics problem situation that allows the users to solve the problem by making informed decisions. This comprehensive understanding is latent in nature; it occurs only when the collection of its prerequisite states is fulfilled. For this reason, this study seeks to identify these prerequisite states in the relevant literature covering information visualization, visual analytics, knowledge discovery, insight problem solving, and cognitive science. The keywords used to identify the documents include "insight", "actionable insight", "actionable information", "actionable knowledge", and "actionable intelligence". Although there are many ways insight has been defined in the literature, these definitions can be classified under categories based on their central idea. This study names the categories as 1) visual-perceptual insight, 2) data insight, 3) knowledge insight, 4) foresight insight, 5) spontaneous insight and, and 6) navigational insight. *Figure 5* shows the six categories of insights commonly found in analytic-related studies and the number of references. Note that each referential document may consist of multiple categories of insight, and thus can be counted more than once across the insight types.

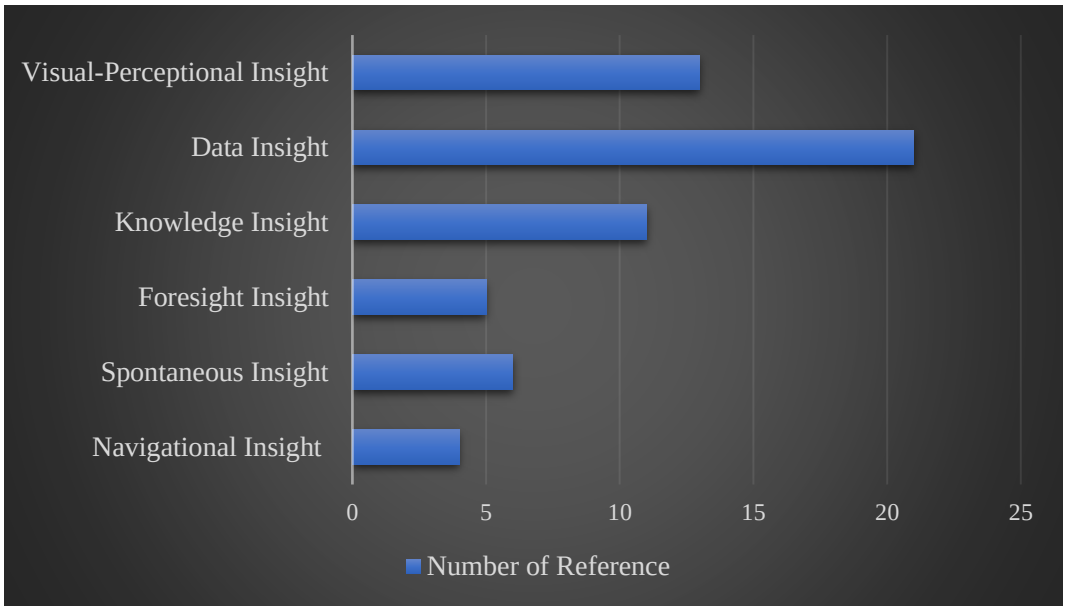


Figure 5. Types of insight in surveyed literature

Visual-Perceptual insight is a specific type of insight that mostly applicable only to the information visualization field. It describes the user perception of visual cues from data visualizations such as graphs, bars, lines, and scatter plots. Visual cues can be perceived from the changes of visual encodings such as color, shape, size, length, orientation, and spatial. This insight focuses on how these external cues are conveyed through human's visual sensory channels. By itself, this insight has very little stake in solving analytics problem because this type of insight is isolated from the data and the context.

Data insight refers to an observation that is relevant to the structure or characteristics of datasets. It has also been known as the "unit of discovery" introduced by Saraiya et al. (2004). Examples of data

insight include the perception of trend, pattern, cluster, detail, or overview. Data insight is often described as the outcome of data exploration. This definition of insight is widely used in a variety of areas, including information visualization, statistical analysis, data mining, business intelligence, and visual analytics. Data insight is gained when users have discovered some key information from the data..

Knowledge insight is known as the change in the way a person understands a problem situation. It occurs when people make connection: 1) between individual pieces of information and 2) between new information in the data and existing mental models in the human mind (Chang, Ziemkiewicz, Green, & Ribarsky, 2009; David & Michelle, 2009). This definition of insight is common for problem-solving in areas such as intelligence analysis and business decision. It has a higher semantic value and less technical information than data insight.

Foresight insight often refers to the prospective information that allows users to project the future states of the problem situation (Watson, 2011). This type of insight is commonly found in more technical research areas such as predictive analysis. Foresight insight is gained when users are able to gauge what is going to happen, based on the coherent collection of knowledge they have gained.

Spontaneous insight is often described as the problem reformulation that leads to the sudden problem solution. Problem reformulation enables new ways of looking at a problem or phenomenon in such a way that its essential features are grasped (Chang et al., 2009). Sponteneous insight is also recognized as an "aha" moment or a "eureka" feeling, when a solution to a problem emerges due to the changes in the perspective the users adopt to see the problem. This definition of insight is commonly found in psychology studies of creative problem solving.

Navigational insight is an intermediate state in iterative and cyclical processes of data analytics. It does not directly lead to the solution, but it informs what the information to look for is, what kind of process is needed, and where the solution could be (David & Michelle, 2009; Gotz, Zhou, & Aggarwal, 2006). This kind of insight is often found in the studies of user interaction, particularly those that involve complex problem-solving.

Based on findings from the literature survey, this study contends that the various insight types exist at different abstraction levels. As shown in *Figure 6*, data insight and visual-perceptual insight have a low abstraction level and are closer to the data, while knowledge insight and foresight insight have a higher abstraction level, and are closer to the semantic aspect of the problem-solving.

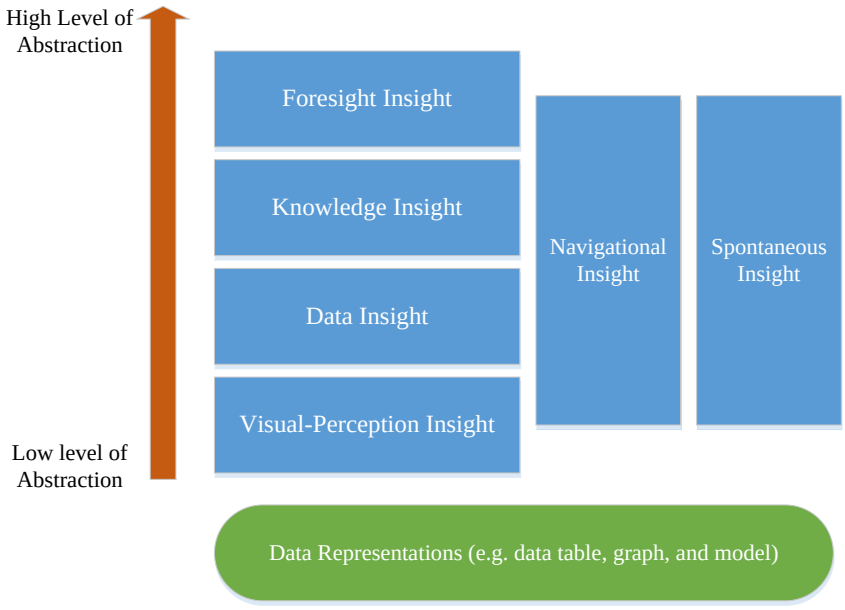


Figure 6. Types of insight and abstraction level

The findings on the understanding of actionable insight suggest two implications. Firstly, there are interdependencies between these insights. For instance, a foresight insight is based on the knowledge insights, whereas a knowledge insight is built on the data insights. Such interdependencies suggest that these insights could be integrated under a unified framework, to provide a holistic view of insight. Secondly, all insight types show that insight is neither data nor information; rather, insight involves reasoning artefact that users have derived from human-information discourse through the usage of the data analytics systems. These two implications, which have provided important clues for this study in the search for relevant theories to explain the insights and corresponding processes to achieve the actionable insight, lead to the two reference theories: situation awareness (SA) theory (see Section 2.3) and sensemaking theory (see Section 2.4).

2.3 Situation Awareness (SA) Theory

From the different types of insight reviewed in Section 2.2, this study suggests that actionable insight is the most comprehensive understanding of the analytics problem that enables the users to solve the problem by making an informed decision. For this latent state to occur, the users need to gain a collection of prerequisite states along the analytical process. More importantly, the review also unveiled that the prerequisite states are conceptually related and can be housed under a coherent framework. Situation awareness (SA) is a theory that provides the basis for such a framework, and that explains the prerequisite states for actionable insight to occur.

2.3.1 Background of Situation Awareness (SA) Theory

In a nutshell, situation awareness refers to the human user's internal conceptualization of a situation. SA was first introduced during the World War I, when military ergonomists began to investigate the

factors affecting air crews, particularly the design of the flight dashboard (Endsley, 1995a). Situation awareness has been recognized from then onwards as a critical foundation for good decision making in complex and dynamic systems such as aviation, air traffic control, manufacturing systems, refineries, and nuclear power plants (Nwiabu, Allison, Holt, Lowit, & Oyeneyin, 2011). The concept was later adopted and established by human factor researchers for studying decision making in complex problem situations (Endsley & Jones, 2011; Shattuck & Miller, 2006). The theory has also been widely used beyond academic research in practical research, particular common among research conducted by Department of Defense in the United States and Australia.

SA theory explains how humans make decisions to take action in real-world settings. The theory focuses on the cognitive states, cognitive artefacts, and cognitive pitfalls of the human decision makers. As opposed to normative decision-making theory, SA theory is well recognized, through reliably explaining more than 80 percent of challenging decisions made by experts in natural settings (Crandall, Klein, & Hoffman, 2006). Endsley (1995b) has also argued that, given that SA theory plays such a critical role in decision making, there is a need to more explicitly incorporate the concept into human-oriented design efforts. In response to Ensley's argument, this study asserts that SA theory can be incorporated into the design of data analytics systems, of which most of the designs largely neglect human factors and behaviors in decision making.

2.3.2 What is Situation Awareness?

Situation awareness (SA) is a set of cognitive artefacts that are collectively defined as "*the perception of the relevant elements, the comprehension of their holistic meaning, the projection of their status in the near future and how various actions will affect the fulfillment of one's goal*" (Endsley, Selcon, Hardiman, & Croft, 1998). Situation awareness is found to be the prerequisite for decision making in complex problem situations (Endsley, 1995b; Li Niu, Jie Lu, & Guangquan Zhang, 2009). To make decisions and take actions, one must first understand the problem situation. For instance, to decide whether to increase sales force in a region, one needs to understand the competitors, customer pools, and local economic conditions. Then the interactions between these factors must be comprehended to get the big picture. Last but not least, the potential impacts of the decision (i.e. to increase sale force) on overall big picture need to be understood.

Ensley's model of situation awareness has three different SA levels, generally known as SA-level 1, SA-level 2, and SA-level 3. The higher levels SA (i.e. SA-level 2 and 3) are particularly critical for effective decision making in complex problem situations. This study asserts that two awareness components can be identified at each level of situation awareness in Endsley's model. For instance, "*the perception of relevant elements*" implies that the *relevant elements* need to be identified before the *perception* can occur. Similarly, "*comprehension of their holistic meaning"* implies the perceptions

need to be *integrated* before the *comprehension* can happen. This study contends that this decomposition provides greater details within the awareness states, and thus, is more useful for understanding how human users behave in complex problem situations. The following three subsections discusses these two components at each situation awareness level.

2.3.3 Situation Awareness Level-1 (SA1)

SA1 is about perceiving the status, attributes, and dynamics of relevant elements in the environment (Endsley & Jones, 2011). The first awareness component at this level is the *identification awareness* that is gained by identifying relevant elements in the complex problem situation. The second awareness component is the *perceptive awareness* that is gained by perceiving the identified elements.

2.3.3.1 Identification Awareness

Identification awareness is gained when users have identified task-relevant elements (Endsley & Jones, 2011; Lu, Niu, & Zhang, 2012). In data analytics, the elements refer to the data elements in the complex analytics problem. To gain this awareness, users need to identify and select relevant data elements from large number of available data elements. For instance, in a task to predict next quarter's sales volume, the relevant data elements for the prediction include previous quarter sales, competitor sales, and market demands.

Identification awareness involves a retrospective process that greatly depends on the users' domain knowledge to determine which data elements are relevant given the current context and objectives. The limited short-term memory and the difficulty of retrieving information from long-term memory impose a cognitive barrier to retrospective thinking (Lee & Chen, 1997). Very often, the users cannot expect all kinds of problem situations and do not always know which data elements are relevant (Ham, 2010). Compounding this issue is the fact that when the complexity of the problem situation increases, users tend to reduce the amount of environmental scanning and focus on familiar yet less information where they had positive results in the past (Parrish, 2008; Thomas, Clark, & Gioia, 1993; Weick, 1995). The overreliance on recallable information from the user knowledge and experience often leads to judgmental error.

As the first awareness component, identification awareness plays a critical role in determining the quality of the other awareness components which built upon it. It acts as an information gatekeeper filtering the data elements that will pass through the problem-solving process. Therefore, for novice users who have less ability to identify meaningful elements, and for users whose elements are limited

to familiar elements only, the effects of inaccurate element identification can propagate throughout the rest of the problem-solving activities (Shattuck & Miller, 2006), thus impairing the quality of the decision.

2.3.3.2 Perceptive Awareness

Perceptive awareness is gained when users have understood the status, attributes, and dynamics of the relevant elements. In data analytics, the data element is only meaningful when being understood in relation to other dimensions such as time and geographical areas. As such, the transformation from data to information is enabled by technical analyses, such as the result from a single regression model or a visual pattern shown on a two-axis scatter plot. At this awareness state, the data has been interpreted into information. For this to happen, the data first has to be transformed into data representations.

Perceptive awareness requires the interpretation that goes beyond simply understanding technical results, to achieving an understanding of the results' significance in the light of the users' objectives, constraints, and context. The interpretation relies heavily on the human knowledge and the heuristics judgment stored as part of the users' mental schemata. The mental schemata provides coherent frameworks for understanding information (Endsley, 1995b). Studies have shown that expert users who have richer mental schemata can notice subtle cues and patterns and make fine discriminations that may not be visible to novice users (Meso, Troutt, & Rudnicka, 2002). In relevant to that, expert users are better and faster at filtering out results that are technically significant but semantically meaningless (Meso et al., 2002). In contrast, novice users might easily get distracted by these results and base their problem solving on these irrelevant results. Endsley and Jones (2011) have also shows other causes of failure to gain perceptive awareness, see Table 2.

Table 2. Major causes of failure to obtain perceptive awareness

Causes	Percentage
Needed information is not provided or is not clear due to system limitations	40%
All information present, but key information was not detected due to display-related issues	33%
Information detected, but was forgotten after the users took in other new information	20%

2.3.4 Situation Awareness Level-2 (SA2)

Situation awareness level 2 (SA2) is about comprehension of the situation as a big picture of the problem (Endsley & Jones, 2011). Endsley and Jones (2011) stated that the comprehension of the problem situation is based on the synthesis of disjointed perceptive awareness from level 1. This implies that the perceptive awareness needs to be synthesized, before the situation comprehension can occur. Likewise,

this study decomposes SA2 into two finer awareness components, namely 1) integrative awareness and 2) comprehensive awareness.

2.3.4.1 Integrative Awareness

Integrative awareness is gained when users have successfully integrated and synthesized the separated perceptive awareness to form a piece of knowledge that is meaningful at the problem-solving level. At this state of awareness, the information has been translated into knowledge. In the data analytics context, the findings from individual technical analyses are integrated to form knowledge that is required to solve the complex analytics problems. The integration can happen in two forms, vertical integration and horizontal integration.

Vertical integration refers to the information synthesis that produces higher-level knowledge. An empirical study has affirmed the importance of synthesizing individual information in an analysis process. The study shows over 80% of users reported that they synthesized two or more separate findings to form a joint conclusion (D. Gotz & Zhou, 2008). Relevant to this, researchers have also pointed out that problem solving is conceptually driven (Lefebvre, 2004). Untrained users do not naturally think in terms of data and the technical relationships between the data. When analyzing a real-world problem to find solutions to the problem, users commonly think, define the problem, and seek for the solution at the concept level. For example, a user wants to find out how political stability and consumer perception toward the company can influence the competitive advantage of a company. Notice that the factors mentioned political stability, consumer perception, and competitive advantage are concepts that humans create and use to represent phenomena or entities in the subsystem of the real world that they attempt to comprehend. Each of these concepts is often the result of the integration of multiple fact-based information items.

Horizontal integration refers to the synthesis between the explicit information within the system and the implicit knowledge that is not stored as part of the system. Endsley (1995b) stated that new information must be combined with existing knowledge to achieve second-level situation awareness. Horizontal integration is introspective information processing where the users are infusing their knowledge, experience, assumptions, and judgment with the factual information in the systems. The result of this macro-cognitive activity is the contextualized knowledge that can be useful for solving the analytics problem. The information discourse between the system and the user's implicit knowledge is important for solving complex analytics problem. However, the process poses high demands on the users' cognitive resources for actively searching for schemata that is relevant to the information in the system. Therefore, the information-knowledge discourse often happens on the fly and is difficult to recall afterward. As the result, the resultant knowledge can be hard to defend and difficult to trace.

2.3.4.2 Comprehensive Awareness

Comprehensive awareness is gained when users have formed and understood a composite picture of the problem (Endsley, 1995b). In the context of data analytics, comprehensive awareness refers to the holistic understanding of the complex analytics problem. As an instance in the stock market, comprehensive awareness involves the understanding of the big picture of how the different factors or entities in the market interact with each other to affect the stock prices. The factors or entities are the higher-level knowledge developed in the previous integrative awareness state.

The previous example implies that one key feature of comprehensive awareness is the interconnectivity between the factors in the complex analytics problem. The interactivity shows how the problem situation works in a closed system. Studies have found that users create a mental replica of the closed system to simplify their problem solving. The creation can happen either subconsciously or consciously in the users' mind. This mental replica is known as a cognitive map. The cognitive map consists of the causal relationships between the factors that are meaningful at the problem-solving level. Studies have found users' capability to create the cognitive map will facilitate better understanding of complex and ill-structured problems.

In other words, a cognitive map is a memory representation expressing the state of affairs in terms of concepts, principles, knowledge, and their relationships (Klein, Moon, & Hoffman, 2006a). The process of developing the cognitive map puts a heavy load on the working memory of users. Due to the limited working memory, the number of factors that users can have in their cognitive space are very limited. One study also stated that, even if users have a very large and complex cognitive map, they may able to access only a small part of it in making a decision (Lee & Chen, 1997). As a result, the cognitive map is a highly subjective, not structurally illustrated, or perhaps messy understanding of the problem situation. Added to the complication is that it is common for users to develop multiple versions of their cognitive maps simultaneously (Kandel, Paepcke, Hellerstein, & Heer, 2012). Without external support, building and maintaining a cognitive map of a problem situation can be cognitively taxing and may reduce the cognitive resources available for the analytical reasoning.

2.3.5 Situation Awareness Level-3 (SA3)

Situation awareness level 3 (SA3) is about the future projection of the problem situation and how various actions will affect the fulfillment of one's goal (Endsley, 1995b). The level of awareness is based on the holistic understanding of the problem situation. SA3 enables the users to be proactive in decision making (Endsley & Jones, 2011). Like the previous awareness levels, this study breaks SA3 down into two separate awareness components, predictive awareness and simulative awareness.

2.3.5.1 Predictive Awareness

Predictive awareness is gained when users are able to predict the future states of the problem situation and the factors within it. In the context of data analytics, the projection involves using various predictive techniques to accurately predict the future states of the analytics problem, based on the existing data. But these predictions happen at a much lower-abstraction level and often are not directly useful for problem solving. For example, predictive techniques can easily predict the future sales volume. However, this prediction often addresses only a small part of the bigger problem the users have, for example, to decide budget allocation to different products lines. In contrast, predictive awareness refers to the users' foresight about how the overall problem landscape may be in the near future. Such overall projection is based on the high-level understanding of current states of the key factors in the problem landscape and how they would influence each other.

By constantly projecting ahead, users will be able to be proactive in developing solutions for their problem situation. As the result, potential courses of action are likely to be formulated after users have successfully gained predictive awareness. The potential courses of action generated are often associated with various hypotheses. This is because different courses of action might develop differently, depending on the fluctuation in the factors. At this state, the users may not have strong confidence about the courses of action (Klein, 1993). Therefore, it is important to generate more alternative courses of action, to avoid premature fixation on too few courses of action. Study has shown that experts are able to generate and consider more alternatives than to novice users can (Jonassen, 2000).

Predictive awareness involves prospective information processing. As per other high-level awareness states, using current comprehensive awareness to form projections requires a very good understanding of the domain and can be very cognitively demanding (Endsley & Jones, 2011). Users may fail to achieve predictive awareness if their cognitive resources are overloaded with other information processing, such as recalling the complex interconnections between the factors in the cognitive map or lacking the domain knowledge. Without external support, few people have the capability of thinking out for more than three months at a time (Primozic et al., 1991).

2.3.5.2 Simulative Awareness

Situation awareness is gained when the users have successfully understood the effects of their courses of action on the future states of problem situation. Simulative awareness is built on top of predictive awareness. Endsley's original model of situation awareness does not contain simulative awareness, neither explicitly nor implicitly. However, given that the purpose of situation awareness is to describe the complete comprehension of the problem situation, this study suggests that, after users are able to forecast what would happen, they are also interested in asking various "what-if" questions to test various

assumptions. More importantly, it is important to get a feel of how their potential solutions would affect the problem situations.

This study draws inferences and support from studies of both recognition-primed and naturalistic decision making to support the idea of simulative awareness. After generating course of action, the decision maker in the natural setting tends to evaluate their courses of action through mental simulation (Klein, 1997; Shattuck & Miller, 2006). They use mental simulation to envision how the solution might work and what obstacles are likely to be encountered. Klein (1993), one of the pioneer researcher in mental simulation, stated that users use mental simulation to test the hypotheses about their potential courses of action under various scenarios. Each scenario is made up by having different values or configurations at each factor in the complex problem situation. The mental simulation process is very cognitively taxing, as it requires the users to mentally maintain the overall structure of the problem situation, and to mentally imagine how the effects are propagated through the structure, based on their understanding about the interrelationships between the factors in the structure.

Given the cognitive resources required by mental simulation, the process often focuses on one change at a time in the problem situation. It involves a deep search of the particular setting to seek out flaws, to find ways around the flaws, and to reject the setting (Klein, 1993). However, the process is not as easy as it sounds theoretically; A study has shown that 90% of their participants failed to accurately simulate a model in their mind, even a simple one (Richmond & Peterson, 2001). Their finding further affirms that humans are generally poor at mentally constructing and simulating a model, without external aids. Another common cause of error is the human tendency to attach increased credibility to the first feasible option, simply because it is the first. This fixation also often causes them to be reluctant to alter the first option even in the face of contrary evidence. Consequently, the course of action selected by the users is often not optimal.

2.3.6 Cognitive Obstacles to achieve SA

The discussions of each level of situation awareness shows that the situation awareness involves a reasoning process that is inseparable from the human users and largely happens internally in the users' mind. The common bottlenecks to achieving situation awareness are often human factors, particularly cognitive resources. The following are three cognitive obstacles that can cause users to achieve situation awareness.

Working memory. Working memory and long-term memory are proven to be critical in the process of achieving situation awareness (Endsley, Bolte, & Jones, 2011). Working memory is heavily in demand, particularly for maintaining the cognitive map and simulating the effects of the potential courses of action of the cognitive map. A large cognitive map often exceeds the memory capacity of average users. This shortage in working memory can be relatively easily alleviated by computer support.

Attention Span. People cannot attend to all information at once. A person's ability to perceive multiple items simultaneously is limited because their attention is finite. This in turn greatly limits the amount of SA an individual can achieve. The limitation in attention span can be easily mitigate using computer-based support.

Mental schemata. Knowledge stored in long-term memory structures known as mental schemata is a critical component in determining the quality of SA. A mental schema is a systematic understanding of how something works in the domain, accumulated through experience and knowledge building. Mental schemata are used to assign meaning to data, to categorize similar concepts together, to establish relationships between factors in the problem situation. Computer-based support may not be able to directly enhance mental schemata, but the support can help to enhance the rigor of its outcome. For instance, the conceptual relationship established based on mental schemata can be validated against the actual data.

2.3.7 Situation Awareness, Decision Making, and Performance

Researchers believe that situation awareness is positively associated with decision making and actual performance (Endsley et al., 2011; Jörn Kohlhammer, May, & Hoffmann, 2009; L. Niu, J. Lu, & G. Zhang, 2009). The basis of their belief is Endsley's study that found the relationships in the aviation domain. Therefore, Endsley's study the relationships should be generalized with caution and needed to be empirically tested if the domain is different. The relationships between situation awareness, decision making, and performance are as shown in *Figure 7*.

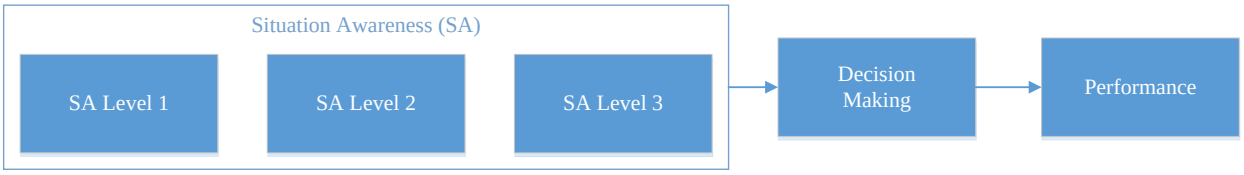


Figure 7. Situation awareness, decision, and performance

It is commonly agreed that higher level and quality situation awareness (SA) increase the probability of good decisions, which in turns lead to actual performance. Hence, studies consider that situation awareness is the key prerequisite for informed decision making. At the same time, situation awareness is also the weakest point in the chain. User studies have proved that more than 85% of human error was occurred in the SA stage (Kokar & Endsley, 2012). That is, their understanding of the problem situation is not accurate.

Given that situation awareness is the key prerequisite for informed decision in a complex problem situation and yet is the most potential to be leveraged to increase the user performance, this study agrees with Endsley et al. (2011) that the most effective way to improve decision is to support users achieving high level of situation awareness along the data analytics process. This call, for a human-driven

approach for designing data analytics, has arisen with the increased realization that system design is no longer optimized for human behaviors in problem solving. Therefore, this study adopts situation awareness theory as the primary justificatory theory for this study.

Assessing users' situation awareness is a key factor in the evaluation of a sociotechnical system (Endsley & Jones, 2011; Salmon et al., 2009). Situation awareness is considered to be an appropriate construct to measure the effectiveness of the system design, because situation awareness is the direct outcome of the information systems. SA reflects the effects of the system design on the computer-human interactions. On the other hand, decision and performance are more of a function of other intertwined factors beyond the information system such as cultural and social pressure, authorities possessed by the decision makers, and the capability of the party who executes the action plan. In short, situation awareness is able to measure the effectiveness of the system design in terms of fulfilling the users' problem-solving objectives, without measuring excessive confounding effects from extraneous factors.

2.3.8 Mapping between Situation Awareness and Insights

The review of situation awareness and the various insight types (see Section 2.2) found that they both have conceptual commonalities. The following *Figure 8* shows how situation awareness and insights are related.

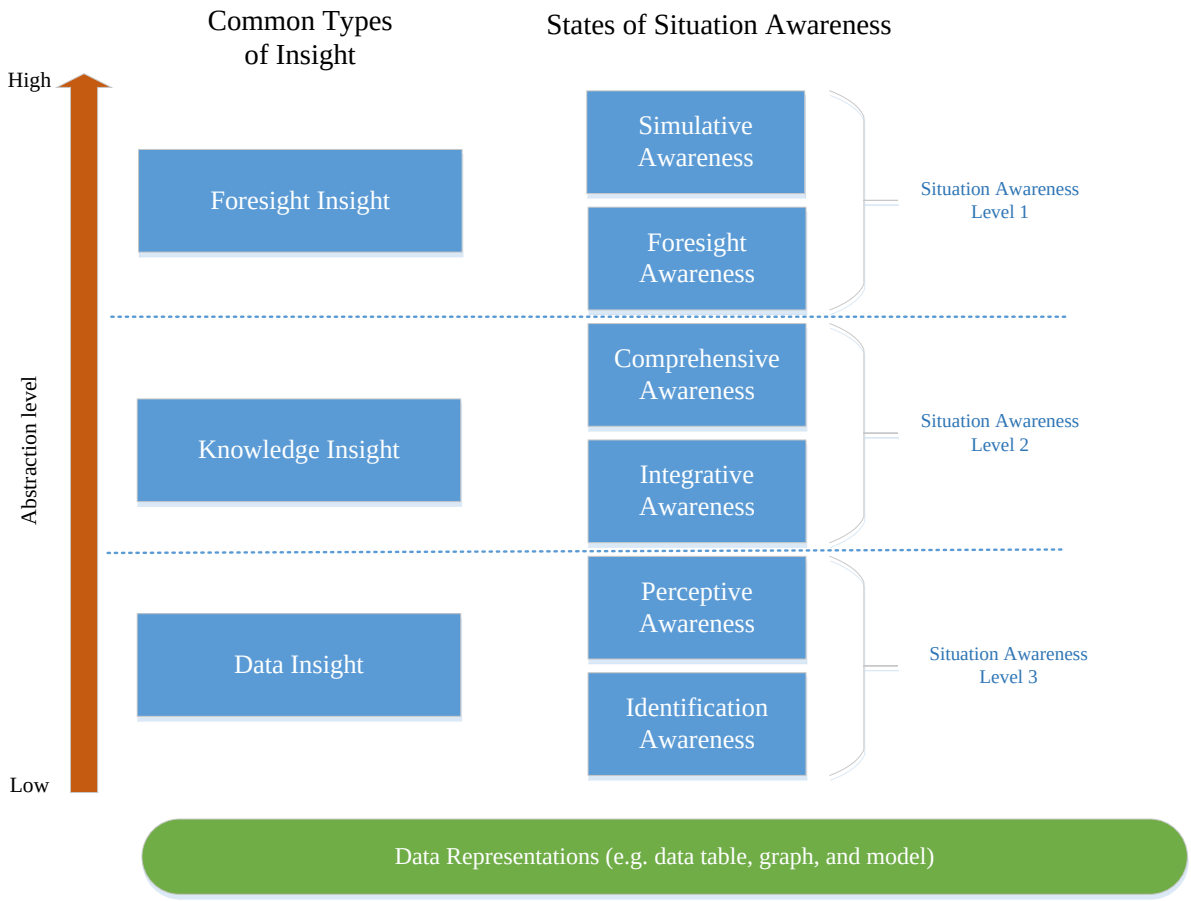


Figure 8. Commonalities between situation awareness and insights

Overall, although the insights are closer to the data analytics context, their available explanations are often limited and relatively abstract. In contrast, the states of situation awareness are more detailed, and thus are more useful in informing the design of data analytics systems. More importantly, based on this finding, this study adopts situation awareness theory to provide a theory-grounded way to organize different types of insight and to explain the relationships between the insights. In other words, situation awareness is used in this study to conceptualize the outcomes of the human-machine-information discourse that occur during the data analytics.

2.4 Sensemaking Theory

Sensemaking theory and situation awareness are closely related. However, the existing literature does not clarify the relationship between SA and sensemaking. This study asserts that, while situation awareness describes the outcomes of the complex problem-solving process, sensemaking theory describes the processes required within the complex problem solving in order for users to achieve situation awareness. In other words, situation awareness provides an "outcome" perspective, whereas sensemaking theory provides a "process" perspective of human behaviors in the complex problem situation.

2.4.1 Background of Sensemaking

Impracticality and rigidness of normative decision-making theory in organizational studies have resulted in a shift toward examining naturalistic decision-making behaviors such as sensemaking (Klein et al., 2006a). Weick developed the sensemaking theory as an alternative approach for understanding how human users process information for decision making in complex problem situations.

By definition, sensemaking is a motivated effort to interpret information in context, to understand connections among information, and to predict future conditions (Klein, Moon, & Hoffman, 2006b; Weick, 1993). The goal of sensemaking is to derive knowledge for actions. Sensemaking theory has a theoretical grounding in symbolic interactionism, ethnomethodology, and sociology of knowledge (Sammon, 2008). Over time, sensemaking has been refined and explicated so that in addition to being a stand-alone theory, it has now started being used as an analysis method in various areas such as business analysis, crime and investigative analysis, and tactical operations (Mills, Thurlow, & Mills, 2010).

Sensemaking theory has long been used in the management field to assist managerial decision making and strategic planning (Parrish, 2008). Instead of focusing on organizational outcomes, sensemaking, a micro-level theory, scrutinizes how an individual uses information to support decisions and actions. More importantly, sensemaking theory provides more representative explanations about domain experts' decision behavior in real-world settings, as opposed to normative decision making theory. The following are the premises of sensemaking theory regarding behaviors in information processing.

- Individual can make decision based on incomplete and uncertain information
- Individual's experience and knowledge play a significant role in the sensemaking process
- Individual uses recognition-primed decision making rapidly to formulate potential solutions
- Individual does not require to exhaustively access all alternative solutions
- Individual's actions are shaped by the ad-hoc and local contingencies of a situation

2.4.2 Sensemaking and Complex Problem Situation

Studies have consistently stated that sensemaking is more useful in complex problem situations (CPS) than in straightforward problem situations (Siemens, 2011; Weick, Sutcliffe, & Obstfeld, 2005; Zhang, Soergel, Klavans, & Oard, 2008). According to Weick's explanation, sensemaking is triggered by uncertainty or an ambiguity problem setting, in which the individual's current knowledge is insufficient to understand and solve the problem. For these complex and uncertain settings, a central

concept of sensemaking is *situation assessment*, a process in which people interpret retrospective and prospective information, in order to develop a situation model that helps them to understand and act upon the problem. From this notion, sensemaking is not only a discovery procedure, but also a creation procedure that aims to drive knowledge for action. Applying sensemaking theory in the data analytics context may help to understand how to design data analytics systems that go beyond mere data discovery to be able to enable actions.

Ability to engage in a sensemaking process has been undeniably recognized as the key to solve complex problem situations (Mirel, 2004; K. Wright, 2005). By generalizing this finding to this current study's context, it implies that sensemaking is the key approach that users adopt to solve complex analytics problems. Thus, the primary role of sensemaking theory in this study is to understand users' behaviors in data analytics from a "process" perspective. The theory informs different stages of analytical activities undertaken by users, and the challenges they face in each distinct stage. Thus, sensemaking theory provides a strong theoretical grounding for this study, from which to identify the leverage points (i.e. design considerations) in which information technology can support the users' processes.

2.4.3 Understanding Sensemaking

The most widely adopted sensemaking process model, developed by Pirolli and Card (2005), consists of two major loops (i.e. stages): 1) the information foraging loop and 2) the sensemaking loop. Each of these loops comprises a set of processes. The foraging loop involves processes aimed at searching, filtering, and collecting relevant information, whereas the sensemaking loop involves creating and testing hypotheses. Pirolli and Card (2005) originally developed their model in the intelligence investigation domain. This study presents the model in the context of data analytics, and has adapted the terms in the model to suite this current context. The sensemaking model is shown in *Figure 9*. The two major loops are indicated as two circular areas; the processes are presented by rectangles.

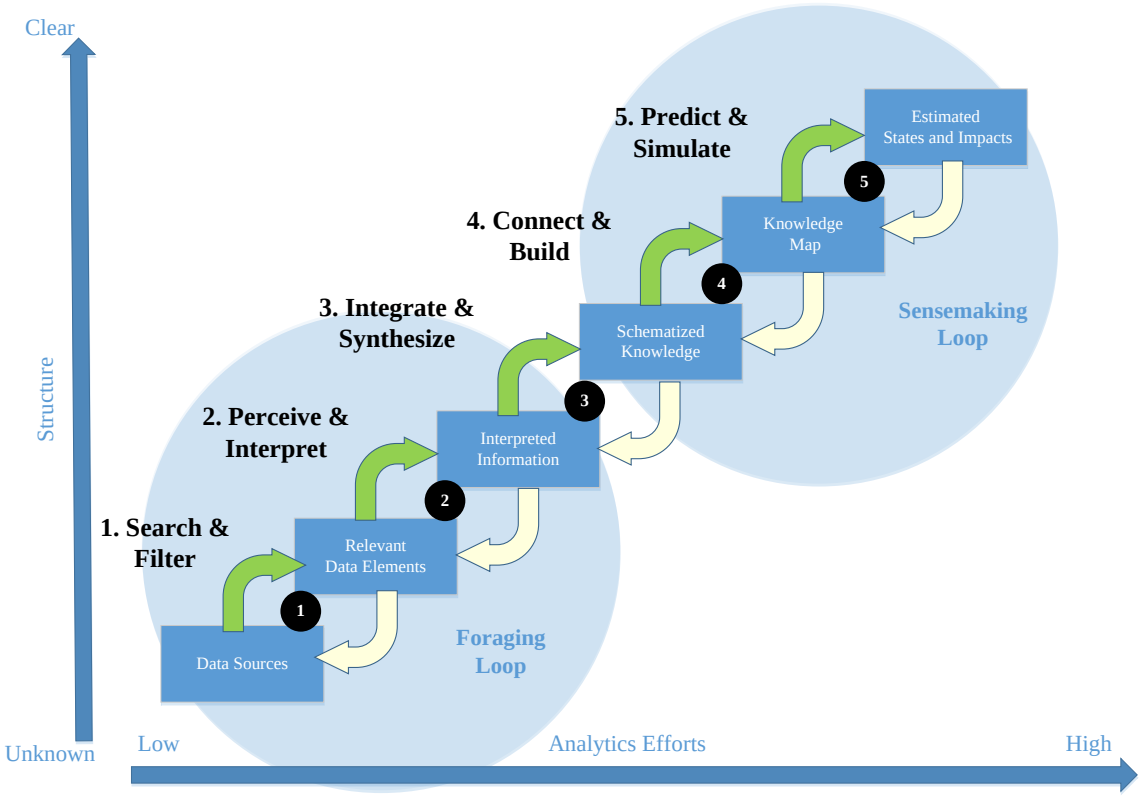


Figure 9. Sensemaking model in business analytics

The sensemaking process can be driven by either a bottom-up (indicated by green arrows) or a top-down (indicated by yellow arrows) approach. The bottom-up approach is based on inductive analysis, where the analyst works from data sources to develop a theorized condition. The top-down approach is based on deductive analysis, where the analyst starts with a theorized condition and finds data or evidence to support the hypothesis. A study on analysts' behavior indicated that top-down and bottom-up approaches are invoked in an opportunistic mix (Pirolli & Card, 2005). Moreover, one may also flexibly switch from one process to another sensemaking process, without carrying them out in a sequential order.

2.4.3.1 Process 1: Search & Filter

This process aims at discovering data elements which are relevant to the analytics problem, from the large amount of data to searching and filtering. Information discovery is a daunting process in complex analytics task because at the early stage of the task analysts often have very few clues about what data could be relevant. At this stage of the analytical task, everything looks attractive to the analysts and the data is often too massive to be explored. The decision maker either may not know what available information relates to the problem or, because of cognitive tunnel vision, may not think to look at pertinent information (Albers, 1999).

The user study conducted by Pirolli and Card (2005) showed that analysts spent significant proportions of their efforts and time on scanning, assessing, and selecting relevant data elements for further attention. The time spent in this process reduces the time they spend on the core analysis processes, thus impoverishing the quality of their analytics outcomes. Studies observed that expert analysts can quickly match the existing problem with their previous experience on similar problems. This allows them to have a quick model in mind to guide their data exploration, to know specifically what information is relevant, and what is not (Gore, Banks, Millward, & Kyriakidou, 2006; Meso et al., 2002).

2.4.3.2 Process 2: Perceive & Interpret

The purpose of this process is to derive information from the relevant data elements. For instance, analysts try to understand a single chart, comparison table, or price trend. Prior to being able to interpret the information, users often need to transform this information into a form that can facilitate the understanding (Jörn Kohlhammer et al., 2009; Russell, Stefik, Pirolli, & Card, 1993). This includes, but is not limited to, techniques such as organizing, grouping, slicing, and summarizing (representing). Representation is critical in manipulating the form of information, which means visualization is helpful for reducing the time and cognitive (perceptual) efforts of users (Albers, 1999). Studies have also reported that visualization is a powerful means of making sense of data (Heer & Shneiderman, 2012b).

It is widely recognized that sensemaking is a domain-dependent process. Simply representing information via a graph, chart, picture, or video, does not necessarily enhance the transmission from data to information (Albers, 1999). Studies stated that the activities in which external representations such as texts, tables, or figures are interpreted into semantic contents are part of the sensemaking process (Celestine A. Ntuen, Park, & Gwang-Myung, 2010). A significant aspect of sensemaking is that it can operate with little or missing data. Study found that users often fill in the gaps through an abduction process, which is a process of forming hunches or reasoning to the best possible explanation (Attfield, Hara, & Wong, 2010). Such unique human ability enables users to derive information from a very small amount of data. Whilst abductive reasoning is powerful, it can easily lead to false interpretations.

2.4.3.3 Process 3: Integrate & Synthesize

The information interpreted in previous processes is often fragmented, and the relationships are obscure. In order to make use of the information they have found, users need to understand the relationships among the pieces, to identify patterns, and to build on their previous knowledge in order to create an updated understanding. The result is schematized knowledge. Schematized knowledge is a semantically meaningful fact-driven knowledge that describes key entities or factors in the problem situation. Schematized knowledge can be formed in two ways. The first involves integrating one or more pieces

of information interpreted in the previous process. The second involves synthesizing the interpreted information with their existing knowledge to form new knowledge.

The schematized knowledge, therefore, implies that it is a composite of coherent information that describes a specific aspect of the problem situation. Pirolli and Card (2005) describe it as the "small-scale story" that answers the who, what, when, where, why, and how questions. In this study, the word "schematized" tries to suggest that the there is a structure organizing or categorizing the information into a rational representation of knowledge. The structure may be created by the users, consciously or subconsciously. Studies have identified very similar behavior, often called "information marshalling", where users gather, organize, categorize, and rearrange information to make it useful for drawing a joint conclusion more easily (Jörn Kohlhammer et al., 2009).

2.4.3.4 Process 4: Connect & Build

Connect & build is the core process in sensemaking. The purpose is to build a big picture representation of the problem situation (Zhang et al., 2008). This process involves continually seeking to understand connections between different knowledge, and then gradually building an "*an architecture of concept relatedness"* (Weick, 1995). Scholars suggest that, at the beginning of the analysis, users bring in their preconceived overall understanding of the problem situation (Yi, Kang, Stasko, & Jacko, 2008), the so called "preliminary frame". Sensemaking, then, is a deliberate, conscious process of fitting fact-driven knowledge into the frame (Celestine A Ntuen, 2009). During the process, the users may update, delete, or discard their frame, as they discover and incorporate new knowledge to change the way they perceive the problem structure. This process is generally known as "mental modeling".

The result is a mental architecture of the problem situation, commonly known as a cognitive map or knowledge map in the area of cognitive study. The cognitive map is a synergetic outcome that takes into considerations the interactions between the individual pieces of knowledge. As such, understanding provided by the cognitive map is larger than the sum of the individual knowledge within it. Siemens (2011) pointed out that the conceptual coherency of the cognitive map is the key for effective decision making and is the premise for confident action. Besides than as a tool for understand a problem situation, the cognitive map can also be used by users to communicate their views and justify their solutions.

The activity to establish the connections between the individual knowledge is greatly relied on the domain knowledge and experience of the users. A study by Meso et al. (2002) shows that experts demonstrate superior ability at framing a problem situation so that the underlying structure of the problem can be detected easily. This is because experience allows experts to build up knowledge templates on which they can draw in new problem situation. Their richer mental models have a lot more declarative knowledge, which can be in the forms of factual statement, rules, associations, and procedures that they can draw on to understand how things work in the domain (Crandall et al., 2006).

As a result, experts are able to apply the declarative knowledge fluidly to a new problem, enabling them to understand a wider range of connections between the key factors or entities in the current problem situation. Moreover, a study by Klein et al. (2006b) in a real-world decision-making setting indicates that skilled decision makers actively build and elaborating competing frames once they detect inaccuracy or inconsistency in their current frame. In other words, the users might develop multiple cognitive maps concurrently.

2.4.3.5 Process 5: Predict & Simulate

Sensemaking has been defined as "*A motivated, continuous effort to understand connections in order to anticipate their trajectories and act effectively* (Klein et al., 2006a)". Predict & simulate process involves 1) using previously understood connections to hypothesize their latest states or to predict the future states and 2) understanding how various actions will affect one's objectives and constraints in the theorized states (Thomas & Cook, 2005). The purpose of sensemaking to derive knowledge for action is hinged on this last process. Predict & simulate process enables the users to gauge the degree and impact of the actions, and thus, be able to proactively allocate resources for making the required change.

Weick (1995), a most influential scholar in sensemaking research, stated that the prediction in a complex problem emphasizes more on plausibility, rather than accuracy, then explained plausibility as the notion that a given theorized state is justifiable, whereas accuracy as the notion that a given theorized stated is better aligned with the facts of a situation than any other understanding. From this understanding, this current study suggests that it is important to improve the users' sense of the possibilities covered by their theorized conditions of the problem situation.

Acting effectively implies optimal allocation of resources. Resource allocation, which in turn, highly depends on how the theorized scenarios may vary and significantly affected by uncertainty. For this purpose, users often mentally experimenting their courses of action with their mental replica of the problem situation's future states. This process is generally known as mental simulation. Due to the uncertainty in the mental replica, the users have various "what if" questions that test out different uncertainties. As the result, multiple versions of the future states are theorized. It is important that the users to be able to understand the effects of these potential scenarios have on their objectives and constrains, so that the resource allocation can be planned optimally.

2.4.4 Mapping between Sensemaking Theory and Situation Awareness

The review of sensemaking theory adds two different perspectives that complement the understanding of complex analytics task, namely the process perspective and the analytical outcome perspective (see Figure 10, two columns on right side).

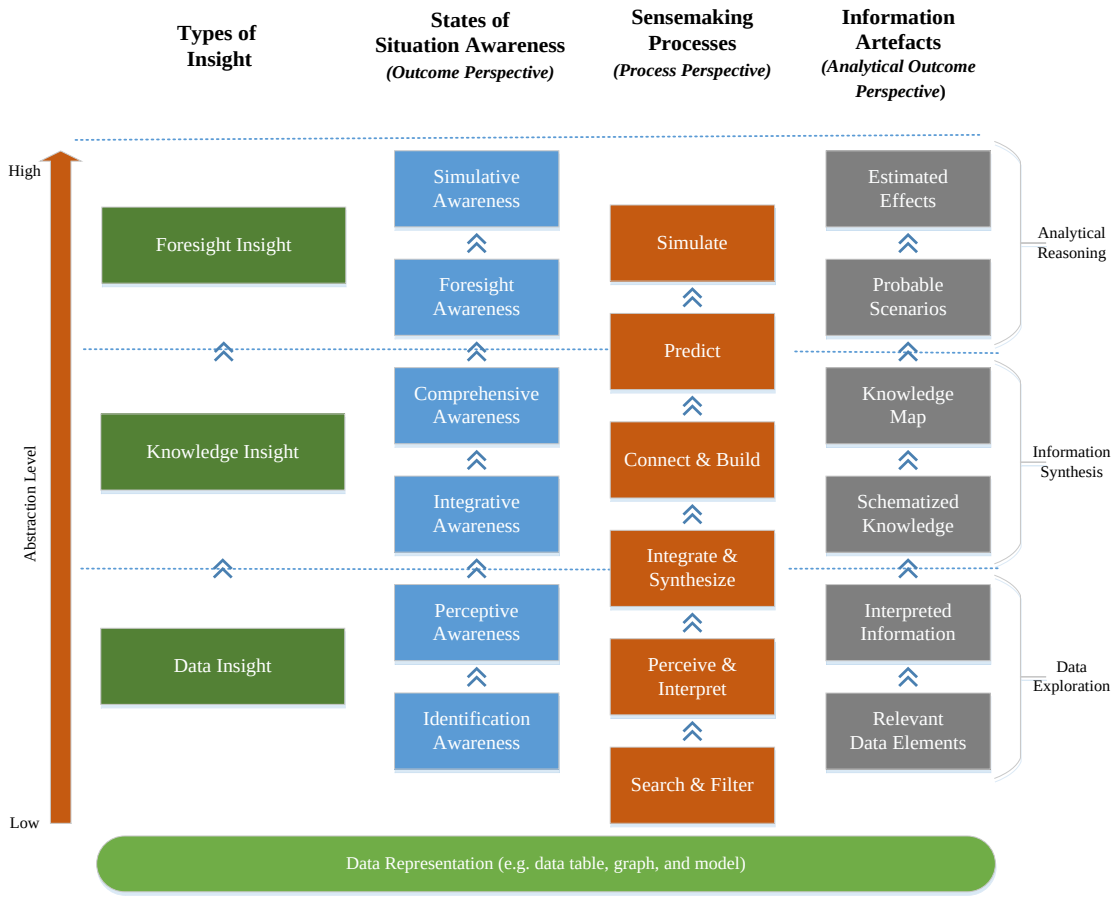


Figure 10. Connections between the sensemaking theory with SA theory

Figure 10 also shows how the two perspectives from sensemaking theory are related to the situation awareness (SA) theory. The process perspective from sensemaking provides the problem-solving activities that correspond to the awareness components in SA theory. For instance, the search & filter activity leads to identification awareness, while the connect & build activity leads to comprehensiveness awareness. On the other hand, the analytical outcome perspective provides a complementary view to the outcome perspective provided by situation awareness. While the situation awareness focuses on cognitive outcomes that are abstract and cognition-oriented, the analytical outcome perspective inferred from sensemaking theory gives more specific descriptions on the outcomes of the complex analytics task. The outcomes are described specifically in terms of information-related artefacts that can be stored and displayed by information systems. This study believes that these information artefacts are intermediate products that are required before the awareness components from the SA theory can be realized in the consciousness of the data analysts.

More importantly, based on this finding, this study adopts the situation awareness theory to provide a theory-grounded way to organize different types of insights and to explain the relationships between the insights. In other words, situation awareness is used in this study to conceptualize the outcomes of the human-machine-information discourse that occurs during the data analytics.

2.5 Complex Problem Situation

The primary objective of users using data analytics systems is to solve a real-world problem. The "*problem*" in this study is a neutral term which refers to the gap between the desired state and the existing state. In this notion, problems can refer to a corporate social responsibility crisis faced by a company or a market expansion by developing new product lines. Right and optimal resources allocation is the key to solving these problems. The resources allocation strategy is in turn relies on the data-driven decision making enabled by data analytics. Therefore, problem solving in complex problem situation is directed toward acquiring actionable insight (Cross & Sproull, 2004)

To effectively design data analytics systems for these problems, it is important to understand the nature of the problems. In practice, most of the problems faced by practitioners such as fund managers, policy makers, and investigators are mostly complex problems. These problems have a set of settings that have significant influence on user behaviors, workflow, and the operating environment in data analytics (Mirel, 2004). Understanding complex problems helps this current study to identify the leverage points where support can be provided, to aid users to counteract the effects of the complex problem.

The complex problem situation (CPS) has long been discussed in the field of cognition science, which examines neurological and psychological aspects of how people think and reason in complex problems. More recently, the concept has become popularized in the area of decision making, especially decisions related to complex socio-techno problems, such as public policy decisions, business and financial decisions, and intelligence analysis. CPS has been an important concept in these problems as they require complex analyses (Keim et al., 2010).

Subsection 2.5.1 introduces the background of the complex problem situation (CPS) concept. Then, subsection 2.5.2 – 2.5.5 presents the characteristics of the complex problem situation by categorizing them into four groups: complexity, interactivity, uncertain, and dynamicity. Each characteristic and its implications, in term of data analytics, are discussed. Subsection 2.5.6 discusses the implications from the understanding of the concept in data analytics.

2.5.1 Characteristics of the Complex Problem Situation

The definitions and explanations of the complex problem situation are scattered across various academic articles and books. In order to have a more complete understanding of the complex problem situation, this study consolidated the piecemeal information from the different sources. Following the widely adopted notion of complex problems from Mirel (2004), this study categorizes the characteristics identified from literature into four dimensions, namely 1) complexity, 2) interconnectivity, 3)

uncertainty, and 4) dynamicity. *Table 3* shows the dimensions and the corresponding characteristics of complex problems.

Table 3. Characteristics of complex problem situation

Dimensions	Dimensions	References
Complexity	Data Volume	(Mirel, 2004) (Klein, 1999) (T., Schreck, Fellner, & Kohlhammer, 2012)
	Data Multidimensionality	(Shattuck & Miller, 2006) (T. et al., 2012)
	Data Heterogeneity	(Shattuck & Miller, 2006) (T. et al., 2012)
Interconnectivity	Interrelated variables	(Funke, 2010; Keim et al., 2010; Mirel, 2004) (Siemens, 2011)
	Synergetic	(William, Richard, & Alan, 2007) (Siemens, 2011)
	Conflicting Objectives	(Crandall et al., 2006; Thomas & Cook, 2005)
Uncertainty	Incomplete Information	(K. Wright, 2005) (Shattuck & Miller, 2006) (Klein, 1999)
	Unpredictable Patterns	(Mirel, 2004) (Sell et al., 2008)
	No Absolute Solution	(Pohl, Smuc, & Mayr, 2012)
Dynamicity	Emergent Process	(Funke, 2010) (Mirel, 2004) (Pohl et al., 2012)
	Nonlinear Process	(Shattuck & Miller, 2006) (Kirsh, 2009)
	Iterative Process	(Heer, Mackinlay, Stolte, & Agrawala, 2008) (Green, Wakkary, Arias, x, & ndez, 2011)

2.5.2 Complexity

Complexity underpins the data setting in complex problems (Mirel, 2004). The complexity of the data can be described by three characteristics: 1) data volume, 2) data dimensions, and 3) data heterogeneity.

Complexity is not only the result of a massive amount of data records, but is also due to the fact that each data record consists of multiple dimensions. For example, stock market analysis is not just concerned with a hundred million rows of historical price data for the companies, but also the few hundred columns of the companies' attributes such as financial ratios, items in financial statement, and performance indicators. Moreover, each of these attributes is associated with other data dimensions, such as time-series. The complexity grows exponentially with the data dimensions. Chabot (2009) stated that even a small number of dimensions can make the data really difficult to analyze and reason.

Although the complex problem solving literature commonly stated that the data in complex problem situation is heterogeneous, often no further explanation is given. By integrating with literature from big data and analytics, this study suggests that data heterogeneity is in terms of the data types such as such as text, raw numbers, statistical formula, computational rules, derived numbers, or multi media. Very

often, a data analytics system is useful for certain types of data. For instance, visual analytics systems are useful for analyzing numerical data, while certain computational analytics systems are useful for extracting patterns and associations in semi-structured data. As implied, multiple isolated data analytics systems are often used in a single analytical task, resulting in islanded results that need to be bridged, manually by the users across the systems, to have a more complete picture of the overall analysis.

2.5.3 Connectivity

Interconnectivity underpins the inner nature of the data in complex problems. The inner nature of the data can be described with three characteristics: 1) interconnectivity, 2) synergy, and 3) conflicting objectives.

Explicit interconnectivity between the data is represented by interweaving the primary key and secondary key relationships in a database. The real challenge is often the implicit interconnectivity that is based on semantic meanings and domain rules that bind the data together, and that may not be represented in the database. Due to the interconnected relationships between the data elements, a change in a single data element can have major effects as the change propagates in a form of effect chains to affect other elements. As a result, even the effect of a simple change can be very difficult to trace and understand.

Complex problem situation is synergic in nature. Given the scale of a complex problem situation, the issue under investigation is too large, and hence requires the users to break down the problem into smaller low-level questions that can be addressed directly by analyzing data. Nevertheless, due to the synergic nature of complex problem situation, adding up these low-level inquiries does not address the overall problem. Useful information typically resides in the overall relationships between the inquiries, rather than in the individual inquiry. In other words, a complex problem situation often requires users to synthesize information from various inquiries in order to derive high-order knowledge that is useful for the overall problem solving (Y. B. Shrinivasan, Gotz, & Jie, 2009; Thomas & Kielman, 2009).

Interrelated objectives and constraints make up a prevalent characteristic of complex problems. The objectives and constrains often conflict with each other (Chiu & Tavella, 2008; Eick, 2000). A great challenge for the users in solving a complex analytics problem is to make choices that simultaneously satisfying both the conflicting objective and the constraints. For example, production's objective to increase product quality conflicts with marketing's objective to strive for competitive pricing. The tradeoff in the objectives and constraints can become complicated due to the interconnectivity of the data elements. The cognitive efforts and information processing capability required in this aspect of the task commonly far exceed the capacity of the users. As a consequence, the users are more inclined to rely on their intuition and heuristic in satisfying the conflicting objectives and constraints.

2.5.4 Uncertainty

Uncertainty is the dimension of complex problems that is associated with the nature of the problem, can be described by three characteristics: 1) incomplete and inconsistent data, 2) unpredictability of the data, and 3) multiple solutions.

A common challenge for users is to deal with incomplete and inconsistent data. Very often, not all the data required is available. Even when it is available, some data can be unreliable and inaccurate. As an example, in stock market analysis, items in the financial statements are often missing, as the items published by different companies can be different. This arbitrary missing data can impair the effectiveness of computational models. Firstly, the missing values reduce the accuracy or goodness of the models in terms of estimation, prediction, or optimization. Secondly, a severe case of missing values may require the models to be modified, which may result in an over-fitting model, which can be misleading. In certain cases, the missing data need to be interpolated based on the available data. The interpolated data can be vastly different because of the techniques selected and the range of data used. As a result, there can be many versions of the data to be considered and these can be unreliable for the purpose of analysis. In short, the data and information in a complex problem situation are subject to human interpretations. The interpretations can be largely varied, depending on the user's understanding of the contextual information and domain knowledge. Users have to fill in the voids of incomplete and uncertain information based on their judgment and knowledge

Another characteristic that contributes to the uncertainty is the unpredictable or nonlinear patterns in the data (Albers, 1999). Data with such nature is a large detriment for conventional data mining and mathematical techniques to be accurate. Mirel (2004) have described its implications as: *imposes the need for variety and complexity of the interpretations that are necessary for deciphering the multiple world-views of the uncertain and unpredictable future*". Besides user knowledge and experience, such data also needs more intelligent and fluid machine learning techniques such as Bayesian network, Markov Chains Network, and neural network modeling to identify meaningful relationships in the data with uncertainties. However, these advanced techniques commonly require the users to have considerable level of relevant knowledge. As the result, such techniques often only benefit a small number of expert users, rather than the general users which are greater in number.

A complex problem situation often has more than one solution (Jonassen, 2000). Due to the open-ended nature of the problem, the solution is highly subjective and there is no right or wrong answer to the complex business problems (Parrish, 2008). More importantly, there is no immediate test for the solution, and any solution may have other consequences for an unbounded period of time (Courtney, 2001). Moreover, the costs of failure are often too large for a trial-and-error approach to the solution.

Added to that, the consequences are nearly impossible to be undone. Environment and political policy making are the best examples that illustrate this issue. As the result, users face great challenges to learn from previous experience. As testing the solutions in the physical world is costly and undesirable, it is critical for the users to be able to assess their solutions virtually with the aids of computer-aided simulation.

2.5.5 Dynamicity

Dynamicity describes the three process characteristics of the complex problem situation: the 1) emergence, 2) nonlinearity, and 3) iteration.

The complex problem situation characterizes an emergence process of which the path to the solution cannot be predetermined (Mirel & Allmendinger, 2004). Scholars have stated that the task to identify and understand the problem is more challenging and important than developing the solution, when dealing with complex problems (Glykas, 2010; Yadav & Khazanchi, 1992). At the initial stage, it is often not clear what the problem exactly is, how a solution might look, and which methods might be used to reach the objectives. The users need to improvise on the analysis strategy and revise their objectives as they proceed. The problem structure is become incrementally clearer as the users progresses through the data analytics. Through the experimental interaction with the problem, users learn more about the problem structure and their analysis strategy's effectiveness. Subsequently, they can modify or refine their interaction to take further analytical actions. The analysts' goal and data interests also tend to change over the course of the process (David Gotz et al., 2010). As a result, automated techniques based on preprogrammed heuristic are often not useful. The process largely relies on the domain knowledge, experience, and heuristic judgment of the human users to progressively seek for the problem solutions.

Nonlinear process is a key feature in the complex problem situation. Scholars have widely agreed that the process is very "messy" and not linear (Heer et al., 2008; Mirel, 2004). Different aspects or smaller components from a complex problem situation are often processed in parallel inquiries. Additionally, studies also show that the users tend to switch quickly from one inquiry to another. For instance, once the users have developed three different stock portfolios, they may start accessing the effects of the portfolios by simultaneously running three separate simulations. Then, the users will attend to the simulation that finished first and proceed in that set of portfolios with further steps before they come back to attend to the other two portfolios. This characteristic of the complex problem implies that the users require the flexibility to switch between different inquiries to support their train of thought, otherwise the flow of reasoning can be disrupted (Green, Ribarsky, & Fisher, 2009).

The iterative process is another common characteristic in a complex problem situation. The problem requires users to access and apply relevant information in iterative cycles to refine the solutions, the

objectives, and the problem structure. A user study recorded that over 96% of participants employed some form of repetitive action patterns within their analysis, strong evidence that the repetitive patterns are indeed a primary structure of user behaviors in complex problems (D. Gotz & Zhou, 2008). Another study has also pointed out that users traverse back and forward from raising the hypothesis to search for supporting information (Pohl et al., 2012). Similarly, Mintzberg and Westley (2001) noted that users tend to constantly respond to events and new sources of information and to work in continual cycles to refine the problem and solution. Parrish (2008) implies that there is no clear stopping rule in a complex problem. As the consequence, users are inclined to make choice that suffice, rather than optimize in complex problem (Jonassen, 2012). Multiple studies have found that users tend to work through iterations of deductive and inductive analysis when dealing with complex problems (Mirel & Allmendinger, 2004). Nevertheless, most data analytics systems often have interactions that are too restrictive to support both inductive and deductive approaches.

2.5.6 Summarized Key Points from Complex Problem Situation

- **Susceptibility to Human Bias -** Users are inclined to rely on intuition and heuristic when the load is beyond their processing capacity. They are also inclined to settle at satisfactory, rather than the optimal solution. Interpretation of the data can largely vary depending on the user's knowledge.

- **Cognition Intensive Process -** The interaction effects of the data elements exert high-level of cognitive loads on the users, in terms of attention, working memory, and reasoning capability. Data dimensions exponentially increase the complexity of the data, making it difficult to reason.

- **Human-Machine Interdependency** - Complex problems rely on human users to perform the actual sensemaking a**n**d analytical reasoning and on the computational power from machine to deal with the complexity and volume of data.

- **Dynamic workflow** – The workflow in a complex problem situation is known to be a nonlinear process and iterative. There can be a mix of inductive and deductive approaches used in one complex problem. Moreover, the users tend to flexibly switch among several concurrent inquiries.

- **Multiple and sometimes conflicting objectives -** The objectives and constrains often conflict with each other. A great challenge for the users in solving a complex analytics problem is to make choices that simultaneously satisfying both the conflicting objective and the constrain

- **Open-structure** – the emergent process involved in complex problems requires users to progressively unfold and learn the problem structure. The understanding of the problem is more important than finding the solutions. Rarely are two inquiries performed in exactly same way, even with the same user. Therefore, each process is unique.

- **Hard to learn from –** potential solutions often cannot be tested because the effects are too costly and irreversible. Moreover, the consequence can be hard to evaluate in practice. Given these conditions, a trial-and-error approach in the real world is not practical.

2.6 Summary: A Big Picture of the Justificatory Theories

The concepts and theories reviewed previously are 1) types of insight 2) complex problem situation, 3) situation awareness, and 4) sensemaking. These separate pieces of justificatory theories originate from different disciplines. In their original form, these theories use very different terminologies from the data analytics community and the discussion may not focus on data analysis. Moreover, these theories exist at different abstract levels.

However, when these theories are being integrated, they are able to provide complementary explanations of the complex analytical task. For instance, the complex problem situation describes the overall operating environment on which the data analytics occurs, situation awareness describes the cognitive states needed to be achieved in order to gain actionable insight, and sensemaking theory is used to infer the problem-solving activities required to achieve those cognitive states.

Moreover, there is no single holistic theory that is specifically developed for the area of data analytics. The information is scattered across different sources. As the result of this theoretical gap, data analytics researchers often have to spend a significant of time to infer design requirements from theories from other domains. This practice can result in redundant and inconsistent efforts. More importantly, there is no native framework or theory in which the results from data analytics studies can be used to reflect and improve upon. The lack of such central repository for the collective knowledge of the data analytics domain may lead to the situation where expensive mistakes are being repeated.

Therefore, this theoretical gap leads to a need for a conceptual framework that can be specifically applied to the context of complex data analytics. The resultant framework can be used by data analytics researchers to understand the user behaviors and requirements in the complex analytics tasks. Developing the conceptual framework will first require the seamless synthesis between the justificatory theories. *Figure 11* shows the integrated view of the different justificatory theories.

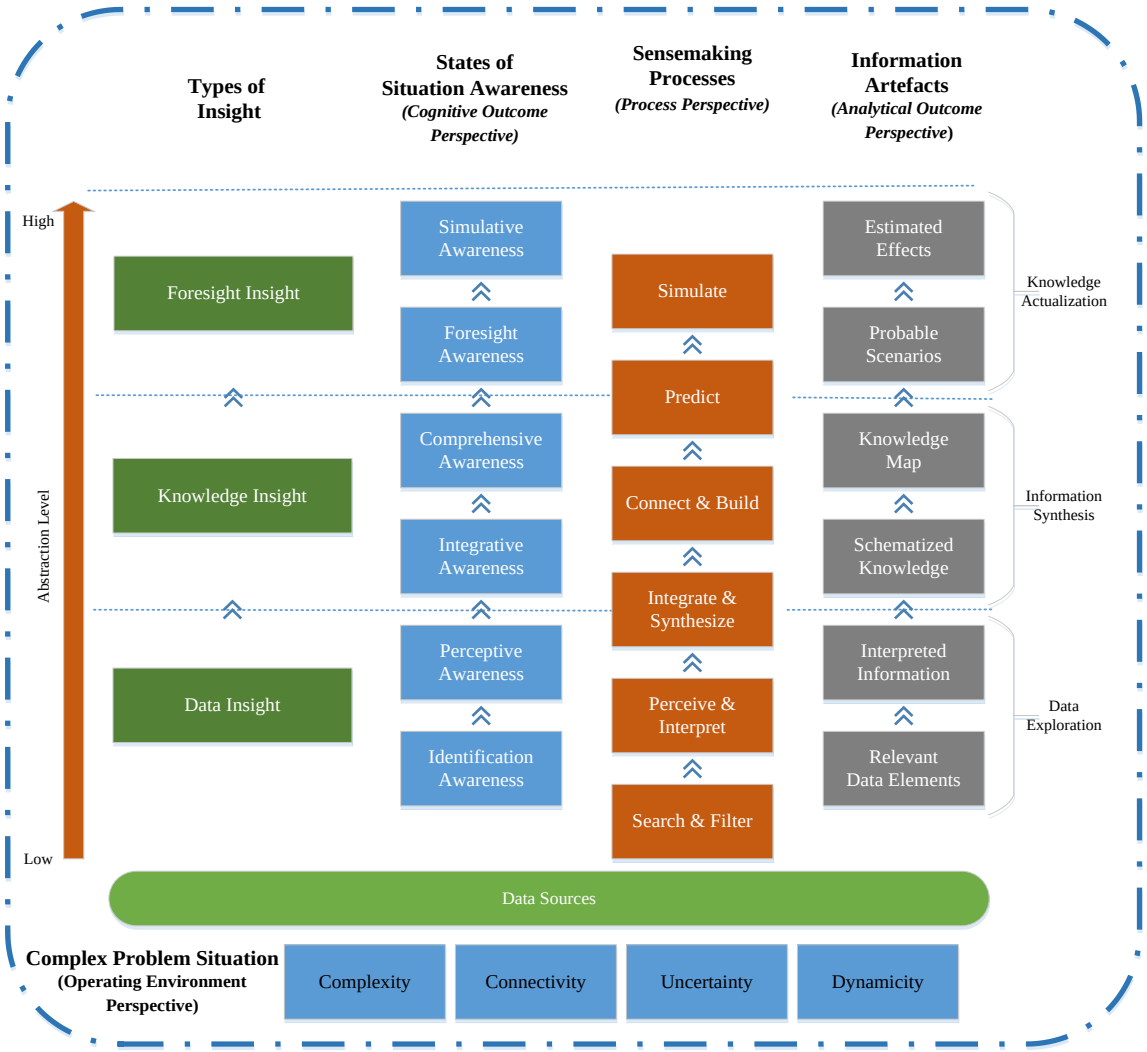


Figure 11. Overview of justificatory knowledge in this study

From the base of *Figure 11*, ***complex problem situation (CPS)*** in **light yellow** describes the overall operating environment in which data analytics tasks are conducted. The concept describes the characteristics of the analytics problem, the data, and the workflow. From the left of the figure are the three ***types of insights*** in **dark green** commonly mentioned in analytical-related studies: data insight, knowledge insight, and foresight insight. Note that navigational and spontaneous insights are not included, as they are beyond the scope of this study; however, the explanations of these insights are often highly abstract and anecdotal, and therefore are insufficient and too vague for informing the designs of data analytics systems.

To mitigate this problem, this study found that ***situation awareness (SA)*** theory in **light blue** can be used to explain these insights in much greater detail, grounding their explanations in a well-established theoretical foundation. For instance, knowledge insight can be broken down and explained by integrative and comprehensive awareness. Situation awareness explains the insights from the perspective of user cognitive states, describing the states of mind that need to be achieved in order to

gain those insights. The theory is also used to explain how various cognitive resources such as working memory, attention span, and metal schemata can limit the user analytical performance. These also reveals the potential leverage points where the system-based supports can be provided to mitigate these cognitive limitations.

Sensemaking theory in **dark orange** provides the process perspective of the data analytics tasks. Precisely, the theory explains the information processing activities required to achieve the awareness states from situation awareness (SA) theory. For example, the *search & filter* activity can lead to identification awareness. The sensemaking theory itself also provides a set of information artefacts that associated with the sensemaking activities. These information artefacts are shown in **grey** on the right side of the figure. While situation awareness theory describes the cognitive outcomes, the information artefacts provide more concrete and detailed information artefacts which can more easily be related to data analytics, and thus are more useful for informing the design of specific system features in the data analytics systems.

From the literature review, this study asserts that a complex analytics task can be divided into 3 phases, namely *data exploration*, *information synthesis*, and *knowledge actualization*, which shown on the right-most brackets in *Figure 11*. Such grouping allows activities to be examined in a way that is well aligned with the processes and outcomes constructs in the aforementioned justificatory knowledge. Note that the **blue dotted line** divided the constructs from the justificatory theories according to the three major phases in data analytics, namely data exploration, information synthesis, and knowledge actualization.

2.7 Related Works

The purpose of this section is to review existing works in the data analytics to find out to what extent existing works are able to support the complex problem-solving activities. Related works from commercial products and research works are also reviewed.

2.7.1 Commercial Products

Visual analytics systems. This type of system aims to support users to explore large amounts of data, by transforming data into visual representations that allow the users to observe and understand the information. The goal of visual analytics is to “gain insight and knowledge”. Studies believe that visual analytics systems can speed up and enhance complex problem-solving by taking advantage of human perception (Eppler & Platts, 2009; Mirel, 2001). This study has reviewed the following visual analytics systems:

- SAS Visual Analytics
- SAP Visual Intelligence
- IBM Cognos Analytics
- Tableau
- Microsoft Power BI

Visual analytics systems rely on the users to generate visualizations, including basic chart types, scatter plots, heat maps, geographical maps, gauges, and parallel coordinate plots. The visualizations allow user interactions such as slice-and-dice, filtering, drill-down, and link-and-brush. Link-and-brush is a technique that allows multiple visualizations to be coordinated to highlight different data dimensions of the same entity, based on user selection. Studies believe that the multi-dimensional data exploration enabled by this technique can help users to glean deeper understanding of the true nature of the data: that is, insight that can be acted upon (Groth & Streefkerk, 2006).

Nevertheless, how visual interactions can help users to gain insight and to solve their problem it is not clearly explained. Some studies have begun to recognize that conducting visual exploration may not be effective for solving a complex problem (Di, Rundensteiner, & Ward, 2007). This current study concludes that most if not all visual analytics systems that are commercially available explicitly support only the first two sensemaking activities, 1) search & filter, and 2) perceive & interpret. In the notion of this study's justificatory knowledge, these systems support users to derive only at data insight, or at situation awareness level 1. In other words, these systems are good at supporting data exploration, but may not be useful for helping the users to understand what they can do with the information discovered. As a result, there is a gap between the low-level data insight and the high-level understanding that is required to make the analysis actionable.

Computational analytics systems. This type of analytics system relies on the computational power of powerful hardware, mathematical models, and intelligent algorithms to extract information from massive amounts of data. A common issue is that the extracted information is still massive due to the fact that the computational techniques human judgment and experience are not taken into consideration in the computation. As Tera Marie Green and Maciejewski (2013) describe computational analytics: "*return the needle in the haystack; however, as the stacks become larger, the problem of producing a needle from a haystack becomes a problem of producing a relevant needle from a stack of needles*". For an instance, association rules mining from a massive set of data can produce a great number of association rules which is still unmanageable by the analysts and resulting a problem on how actionable and meaningful rules can be extracted from these association rules. As the result, scholars believe that a computational analytics system can exacerbate a complex problem, rather than aid its solution (Cao, 2012; Endsley & Garland, 2000).

As with the visual analytics systems, there is a gap here between the information extracted and the high-level understandings required by the users to solve the complex analytics problem. Therefore, this study asserts that most existing computational analytics systems explicitly support only up to data exploration. They do not explicitly help the users to form high-level understandings about the key entities in the problem situation nor hypotheses of which courses of action can be useful to solve the problem. Note that machine learning and data mining may able to produce high-level rules, such as a collection of customers who are likely to churn. However, very often, these information is only a part of the bigger decision context. For instance, the decision makers still require information from other data analysis, such as performance of sales representatives, and effectiveness of promotion efforts, in order to find a solution to improve the sales revenue in the coming quarter. The integration and analysis of such information is seldom supported directly the analytics systems.

2.7.2 Research Works

This study turns to review research works because these works often have more experimental works, prototypes, and early designs. This led this study to discover ARUVI and SRS, which similarly intend to support users beyond data exploration, to higher-level sensemaking activities such as integrate & synthesize information, and connect & build knowledge map.

2.7.2.1 ARUVI

ARUVI is a data analytics system for analyzing multidimensional data via interactive visualization (Yedendra B. Shrinivasan & Wijk, 2008). The system contains three different views, namely *data view*, *navigation view*, and *knowledge view*, as shown in *Figure 12*. The data view is a visual analytics-based tool that allows users to create interactive visualizations, while the navigation view provides an overview of the data exploration process by storing the visualizations. The *knowledge view* is the key point of the discussion, as its purpose is to allow the analyst to combine individual findings from the visualizations into a knowledge map. The authors describe the ellipse shape as the "concept" and the rectangle boxes as the "notes" which can be used to record the data insight gained. Users can also categorize various "notes" within the same group, or can create nested groups. Links between the "notes" can be used to represent different relationships such as parent-child relationship and causal relationships between the "notes".

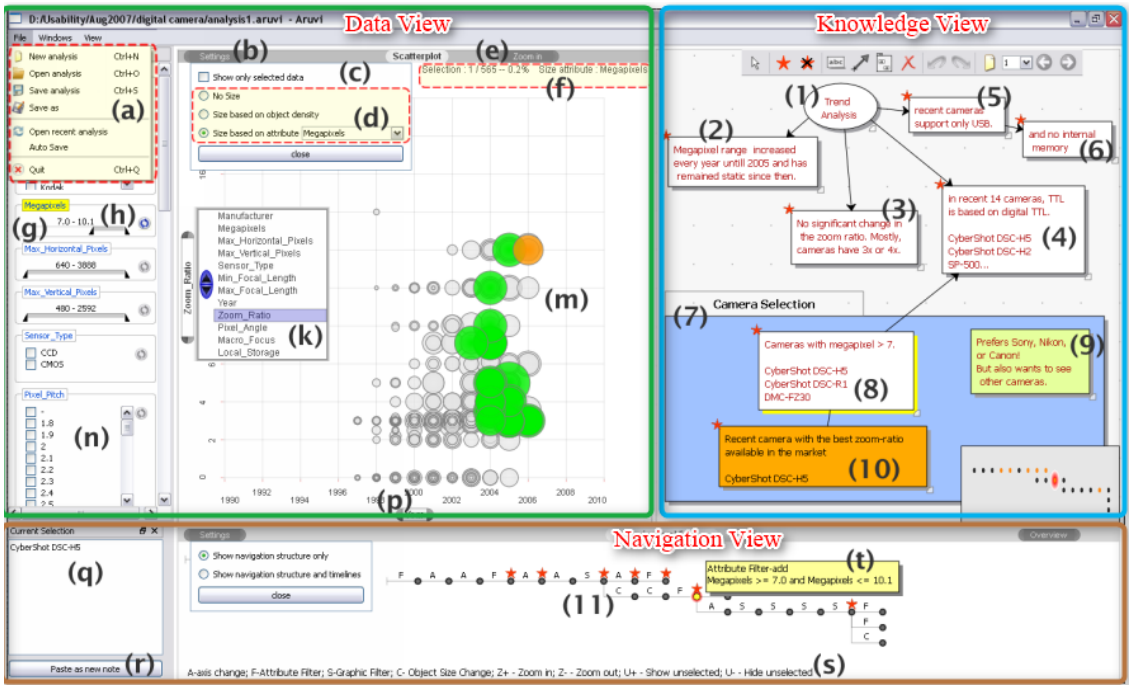


Figure 12. Three views in ARUVI: data view, navigation view, and knowledge view

The strength of the mechanism is that the elements are linked with the respective visualization results. This feature allows analysts to easily retrieve the visualizations that underpin the particular "note" in the knowledge view. Moreover, the nested groups of notes can be used to represent multi-abstraction knowledge. In other words, ARUVI might able to support the "integrate & synthesize" and "connect and build" of sensemaking process.

This system has drawbacks. The "knowledge view" is not more than a mind map diagramming tool that is included as part of the application; it is still far from being an integrated part of the system that works seamlessly with the rest of the modules. This study contends that the "note" mechanism is too simple and inadequate to accommodate a complex problem situation. For instance, there is no operator to represent possibility and the magnitude of influence. Additionally, one of the drawbacks is the need of users to manually enter what they learnt into the "notes" in text form. When there are large numbers of "notes" and a congested screen, the limited users' attention span may limit their access to the knowledge in the "notes" and thus impair their reasoning process. More importantly, given that the knowledge is encoded in the text-based notes, the reasoning relies entirely on the human users, while the system is responsible only for "holding" and "displaying" the knowledge map. In other words, it cannot take advantage of computer-aided reasoning techniques to enhance the reasoning capability of the users. As a result, the human-intensive reasoning process might impair the system's scalability for large and complex problem.

2.7.2.2 SRS

The scalable reasoning system (SRS) is an analytics system that aims to provide a tool for the collecting, analysis, and dissemination of reasoning products (W. A. Pike et al., 2007). Besides "integrate & synthesize" and "connect & build" which generally known as information synthesis, SRS also partially supports "predict & simulate" of the sensemaking process.

Figure 13 shows the interface for information synthesis. Based on the SRS's notion, text, image, multimedia, and hyperlink are *information source* which the users can extract *evidence(s)* from. Subsequently, *evidence* can be combined with other reasoning products such as *assumption* and *argument* to form higher-order concepts such as the drug resistance and the H5N1 pandemic in the figure. SRS provides a more structured way than ARUVI to organize information because each piece of information is categorized to a specific type and represented as a graphical icon that can be easily recognized.

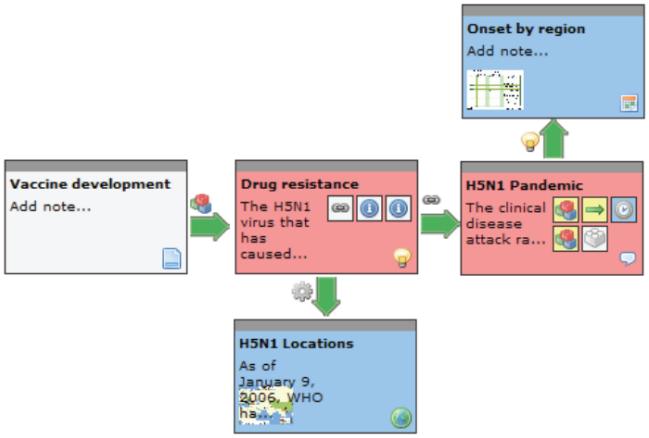


Figure 13. SRS's interface for information synthesis

---- This space is intentionally left blank ----

The true strength of SRS is its capability to support users in carrying out "predict & simulative" activities to assess their hypothesis. SRS depicts reasoning artifacts in the form of a hypothesis network. As shown in *Figure 14*, each of the artifacts can be assigned with confidence value that indicate the extent to which the artifact is believed to be true or false, whereas the links between artifacts can take on probative force values that indicate the magnitude of the relationship. These numerical values can be used to perform mathematical reasoning over a hypothesis network to determine the likelihood of a line of reasoning and analysis of competing hypotheses (W. A. Pike et al., 2007).

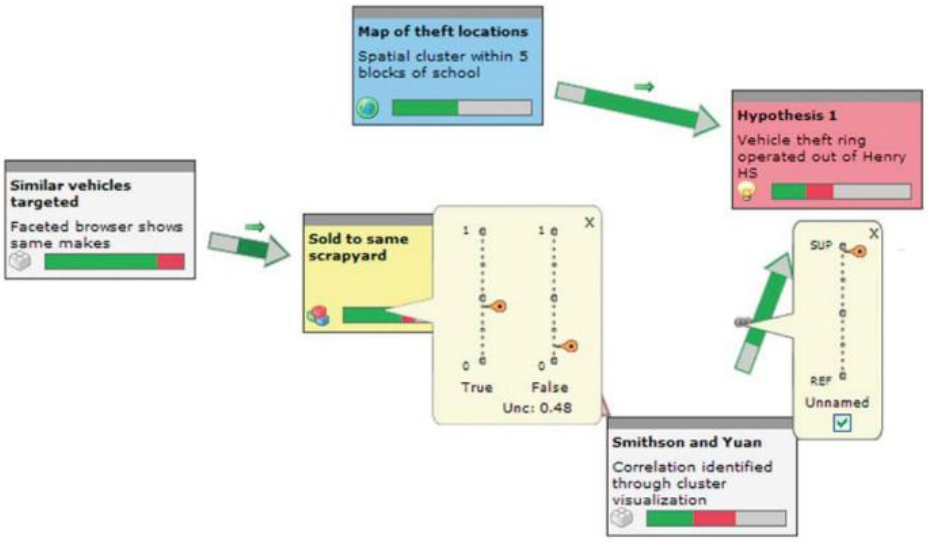


Figure 14. Analysis of hypothesis network in SRS

Although SRS has the basis for being a "working model", the use of mathematical reasoning methods to execute the model is still in the conceptual stage and much work needs to be done. More importantly, SRS is still rather limited in supporting complex analytics reasoning such as comparing scenario, simulating what-if conditions, and incorporating objective and constraints. More importantly, it still relies entirely on the human users to manually make inference from the data and to determine the input values for the modeling. Keim et al. (2010) argue that this user-reliance approach gives good results for small datasets; however, it fails when the data for solving the problem is too large. In addition, this approach is largely subjective, therefore, diminishing the value of the quantitative data collected.

2.7.3 Summary of Review on Related Works

This study has also reviewed works from W. Wright, Schroh, Proulx, Skaburskis, and Cort (2006), David Gotz et al. (2010), and Robinson (2008). Together with SRS and ARUVI, these works made to support users beyond data exploration can be further improved from the following aspects:

- These systems require users to manually annotate their discoveries. Manual annotation is time-consuming, non-scalable, and tends to distract the users from the flow of analysis. Moreover, free-text annotation can be imprecise and hard to understand, with the added difficulty of needing to recall the logics that underpin the discovery.

- Most of the synthesized information is static and non-computational (e.g. chunks of texts to describe the findings). The non-structured information makes the further reasoning to rely entirely on human judgment and reasoning. This characteristic makes the synthesized information less useful for further processing, as it is not suitable for use as building blocks for building structural arguments or hypotheses that can take advantage of structural reasoning methods to achieve rigorous analysis.

- The primary purposes of the information synthesis in these systems are for ease-of-retrieval, traceability to source, and communication. These mechanisms were not designed to support the problem-solving activities of the users sufficiently. As a result, they are lacking of the features for supporting the users to achieve the situation awareness required to solve the analytics problem.

- The features in these working systems are made to "hold" the information for the users, so the users can process the information. In other words, the features focus on alleviating the users from the "attention span" and "working memory" constraints, but do not enhance the analytical reasoning of the users. Therefore, the systems were not designed to improve users' analytical performance.

- With the exception of ARUVI, all other systems were designed for processing qualitative information and are not able to take advantage of quantitative data. However, even ill-structured problems in practice would have quantitative data that can be used to aid the analysis.

- None of the systems are able to support the complete sensemaking process and situation awareness states which are critical for effective problem solving. Most of these research works support up to information synthesis only, while SRS is the only system that is designed to support beyond information synthesis.

- These systems were not designed to specifically handle complex analytical tasks. For instance, there is no feature provided to deal with a highly uncertain and massive dataset. Furthermore, most of these systems rely greatly on humans for the information processing task. Such an approach makes the systems difficult to scale up to a complex problem situation with large quantities of data.

Several major differences distinguish these previous works and this study.

- The study aims to explicit support users on all the problem-solving activities required to solve an analytics problem. In other words, the proposed system was designed to support all the sensemaking activities and the situation awareness states.

- The study aims to target complex analytical tasks in which the data is massive, interconnected, and highly uncertain. The features in the proposed system were specifically designed to deal with that type of problem. This also implies that the system is able to take advantage of quantitative data.

- The purpose of information synthesis in this study is to enable users to see the big picture of the problem as a part of the endeavor to solve their analytical goal. The synthesized information must be computable and dynamic in nature. In short, this study envisages a structured analytical reasoning approach that is not only externally represented using computer memory, but also capable of reaping the processing power of the computer to augment the human reasoning process.

- This study seeks a balance between human-driven reasoning and machine-driven efficiency. For instance, the reasoning loads will not entirely fall on the human users. The loads are divided between the human users and the computations, according to their strengths. This study believes that this human-machine symbiosis approach is capable of reducing human biases, while allowing human judgment to drive the semantically meaningful computations. The result would be a rigorous analytical outcome with high acceptancy from the users.

2.8 Summary of Literature Review

Through the reviews of both commercial and academic works, this study found that most of the commercial products and a majority of the academic projects have focused on the data exploration phase of data analytics. Only a handful of research projects aim to support users at the information synthesis phase and beyond. *Figure 12* shows the support from commercial and research.

In the data exploration stage, users transform the data into information in order to answer what, when, how much, and possibly why questions. Both software support and research at the analytic insight layer are comprehensive and mature. However, most of the works have focused on the technical-driven advancement.

The information synthesis phase involves activities such as abstracting, organizing, assigning semantic meanings to analytics result, and producing new

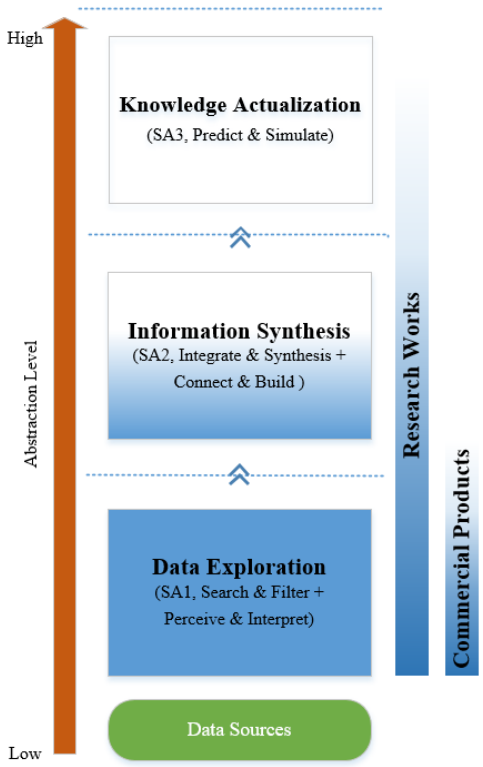


Figure 15. Phases in data analytics and supports

knowledge, based on joint findings (Robinson, 2009). The purpose of information synthesis is to connect those fragmented analytical results into a big picture that reflects the overall analytical task. The lack of support for information synthesis is a well-recognized gap in the field of analytics (D. Gotz & Zhou, 2008; Robinson, 2009; Yedendra B. Shrinivasan & Wijk, 2008). Based on the best knowledge of this study, there is no commercial analytical systems that explicitly support this stage, and there are only a few research projects aiming to do so. ARUVI and Sandbox are the examples. Conceptual research at the synergic insight level has been long existed. However, there is a real need for developing techniques that can realize the conceptual designs. The efforts should go beyond the conventional "evidence marshalling" or "evidence shoebox" which relies on human users for connecting, organizing, and reasoning (Thomas & Cook, 2005).

The knowledge actualization stage is central to the analytical task of applying human judgment to reach conclusions or devise solutions (Ribarsky, Fisher, & Pottenger, 2009; Yedendra B. Shrinivasan & Wijk, 2008). This is the stage which requires most research work and software support. Most of the prediction and simulation techniques are meant for quantitative data and are poor in dealing with uncertain, dynamic, and subjective data.

Without dedicated support, information synthesis and knowledge actualization often occurred outside the data analytics environment, either inside the user's mind or on the paper. Without the proper support from information systems, it is difficult to perform these activities effective and accurately. In notion of sensemaking theory, existing systems focuses heavily on the foraging loop, while neglect the sensemaking loop. As a result, they often fail to support users achieve situation awareness level 2 and 3 that are crucial for making informed decision. Therefore, this study believes that in order to address the problem of which existing data analytics systems fail to deliver action insight, the data analytics systems must be improved to support the complete data analytics phases, ranging from data exploration to information synthesis and knowledge actualization.

---- This space is intentionally left blank ----

Chapter 3
Research Methodology & Design

3.1 Overview

The purpose of this chapter is to present the research methods used in this study. *Subsection 3.2* describes the central research approach that is used to provide a systematic research framework to ensure rigor procedures are taken to answer the research questions. The research design in *Subsection 3.3* describes the characteristics of this study's research method from the aspects of purpose, research setting, time, unit of analysis, and types of the analysis. This information provides the basis for designing and organizing the flow of research activities, which is depicted in *Subsection 3.4*.

3.2 Design Science Research as the Central Research Methodology

This study adopts design science research as the central research methodology. Design science research is a pragmatic approach to solving real-world problems by designing and creating IT artefacts. The methodology research is aligned with the objectives of this study 1) to understand the user behaviors in an IT usage setting, 2) to develop a design to support the user behaviors, and 3) to evaluate the effectiveness of the design. For this reason, design science research is an appropriate methodology for providing a systematic framework for the research outcomes and research processes of this study.

The primary outcome of design research is a *design theory*. Design theory, a cohesive set of design knowledge, is expressed as a theory focusing on "how to do something". Several studies have attempted to specify and formalize the components of a design theory (Arazy, Kumar, & Shapira, 2010; Kuechler & Vaishnavi, 2012; Walls, Widmeyer, & El Sawy, 1992). A more comprehensive list of design theory's components has been introduced by Gregor and Jones (2007) in their article titled "*The anatomy of a design theory*". The research outcomes in this study are influenced by the notion of design theory from Gregor and Jones (2007). The resultant design theory consists of design principles, testable propositions, justificatory knowledge, and instantiations as the components of the design theory. *Table 4* describes the components of the design theory in this study.

Table 4. Components of a design theory

Component	Description
Purpose and scope	• Purpose of the design • Scope, limitations of the design
Justificatory knowledge	• Descriptive knowledge or theories that give a basis for understanding the phenomena or behaviors of interest • Often are from other disciplines or contexts

Constructs	• Factors or variables of interest in the design • Key elements in the behaviors or phenomena of interest • May be manipulated to influence the behaviors
Conceptual explanatory framework	• Explanatory knowledge describing the user behaviors and interaction effects, and hence, lead to the design requirements • Result of synthesizing justificatory theories • A conceptual intermediary between the foreign justificatory theories and the present study's context • Specifies the design requirements
Conceptual design framework	• Abstract architecture of the design • Design-oriented knowledge which explaining why the design has the effects it does • Specifies specific design objectives
Design principles	• Components of the conceptual design framework • Set of design guidelines to achieve the specific design objectives • Operationalize the abstract designs into system features
Design instantiation	• Physical implementation of the design theory which can assist in representing the theory both as an expository device and for the purpose of testing
Testable propositions	• Propositional statements about the design that are intended to be confirmed by testing the design instantiation

Besides the research outcomes, the methodology also provides systematic guidelines for the research process. Design science research often starts with the use of justificatory knowledge to derive design requirements, which in turn are used to develop the design theory (Markus, Majchrzak, & Gasser, 2002; Walls, Widmeyer, & El Sawy, 1992). Subsequently, the design is evaluated to see whether the design achieves its design objectives. *Figure 13* illustrates the general flow of research processes in design science research. This study's research activities, described in section 3.4, are built on this process flow, which is widely adopted among the design science research community.

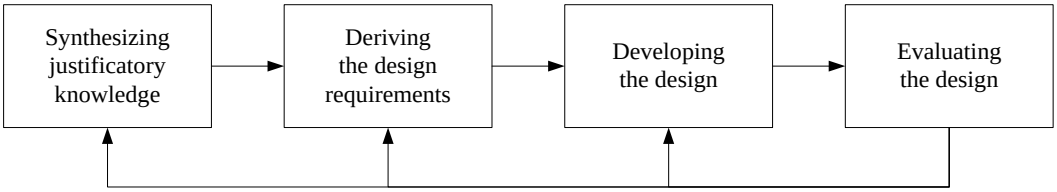


Figure 16 Flow of research activities in design research

As a design science research, this study aims to contribute to the body of academic knowledge and to the practitioners. As the practical contribution, the design theory provides prescriptive guidelines to practitioners in practice for solving problems of the same class. As the theoretical contribution, the design theory provides a tested framework for understanding the phenomenon being studied.

3.3 Research Design

A research design is a set of research aspects that together define the logical structure of a research inquiry. In considering the research questions and objectives, research design seeks to specify what type of evidence is needed and what type of analysis method is appropriate for that particular study. The research procedures and methods can then be determined according to the research design. Research design can be seen as the requirements of a particular study, which contextualizes the general research methodology into a specific work plan.

According to Sekaran and Bougie (2009), the purpose of a study can be exploratory, descriptive, or hypothesis testing in nature. The purpose of this study is best described as hypothesis testing, which seeks to evaluate the effects of the design on users' analytical performance. The effects are observed by comparing the difference between the proposed system and an alternative system. This also implies that this is a confirmatory research which tests a priori hypothesis. Such a priori hypothesis is made before the measurement phase starts and is usually informed by a theory.

A user study was used in this study to collect the data required for the hypothesis testing. The user study is conducted in a controlled environment where the tasks are predesigned and datasets are controlled. More details about the user study are given in Chapter 6. In terms of time setting, this study is cross-sectional in nature, in which the data is gathered once to present the state of a single point in time. The unit of analysis of this study is individual: each participant is treated as an individual data source (Sekaran & Bougie, 2009). Both quantitative and qualitative data analysis will be used complementarily, to ensure that both structured and non-structured evidence is collected to support the hypotheses testing. *Table 5* summarizes the research design of this study.

Table 5. Summary of research design

Aspect of Research Design	Description
Purpose	Hypothesis testing
Nature	Confirmatory study
Main relationships of interest	Difference comparison
Data collection method	User Study
Unit of Analysis	Individual
Nature of Analysis	Both quantitative and qualitative
Time	Cross-sectional

3.4 Research Activities and Flow

This section describes which research activities are required and how these research activities are organized to achieve the research objectives of this study. The choice and flow of the activities are informed by the given research design and guided by design research methodology. This study comprises the following six research phases:

1) Research problem identification
2) Literature review
3) Conceptual explanatory framework development
4) Conceptual design framework development
5) Design Evaluation
6) Design Theorization

Figure 17 shows the specific research activities (left column) and the outcomes of each of the six phases (right column). Note that the outcomes from the research activities correspond to the components of design theory.

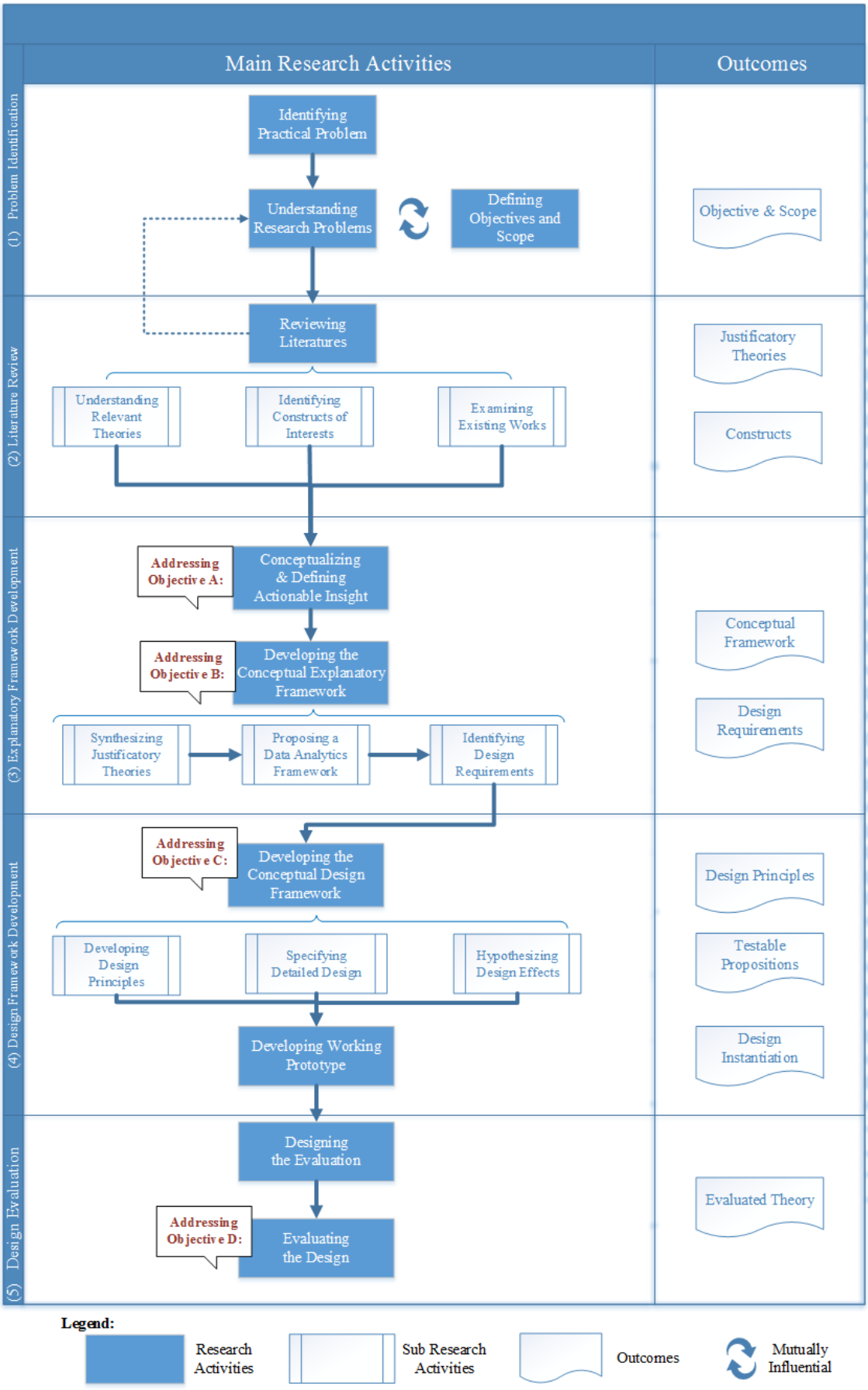


Figure 17 Research activities and flow

The following *Table 6* provides the details of each research activity, corresponding to the flow chart (*Figure 17*).

Table 6. Details of research activities

	Research Activities	Descriptions
(1) Problem Identification	Identifying practical problem	The research idea of this study is informed by real-world problems. Multiple sources, ranging from practitioners, industry reports and magazines technical forums, and academic literature were converged to confirm that the practical problem is a common issue in the domain, rather than a few isolated cases. This activity is to ensure that the contributions of this research are important and practical. Corresponding content: Chapter 1
	Understanding research problem(s)	Extensive literature is reviewed to identify the underlying research problem, instead of the mere symptoms of the practical problem. Moreover, both previous works and commercial products are examined to ensure the originality of this study. This activity was iterated to progressively refine the research problems. This study initially found a number of root problems which could potentially contribute to the practical problem. Then, these problems were prioritized and consolidated in accordance to their impacts on the practical problem and the extent to which they are backed by scientific evidence. This activity is to ensure that this study is built on a strong, yet manageable set of research problems. Corresponding content: Chapter 1
	Defining objective and scope	The research objectives and scope of this study are informed by the research problems. The objective shapes the rest of the research activities in this study. Additionally, the objective and scope imply the overall design goal and boundary of the design theory developed.

		Corresponding content: Chapter 1
(2) Literature Review	Reviewing Literature	This activity serves two purposes. Firstly, it identifies theoretical gaps in the existing literature. Secondly, it identifies justificatory theories for explaining the user behaviors from relevant domains such as cognitive science, problem solving, and decision making. The activity consists of three sub-activities, 1) understanding relevant theories, 2) identifying constructs of interests, and 3) exploring state-of-the-art data analytics solutions in both academia and industry. Corresponding content: Chapter 2
(3) Explanatory Framework Development	Defining and conceptualizing actionable insight	To define the term "actionable insight" to reflect the way this study conceptualizes the term. In this study, actionable insight is conceptualized as a multi-component concept. This research activity aims to address ***Research Objective A***, which is to propose a systematic and theory-driven definition of actionable insight. Corresponding content: Chapter 4
	Developing conceptual explanatory framework	The purpose is to develop a native IS framework to describe and explain the user behaviors, interaction effects, and other phenomena of interest, specifically in the data analytics domain. This activity contains the three forthcoming sub-activities. This research activity aims to address ***Research Objective B***, which is to develop a conceptual framework for understanding the processes and requirements of actionable insight Corresponding content: Chapter 4
	Synthesizing justificatory theories	To develop the framework, this study synthesizes the theories, including situation awareness, sensemaking, complex problem solving, and mental model, in a synergic manner to explain the user behaviors in data analytics.

		This study uses theories that are well-validated and with high creditability to ensure this study's understanding of the phenomena in data analytics is grounded in strong theoretical basis. Corresponding content: Chapter 4
	Proposing a Data Analytics Framework	As the result of theories synthesis, an explanatory framework of data analytics is developed. The framework explains the user states, activities required, information processing, and the leverage points where the user analytical performance can be enhanced. Corresponding content: Chapter 4
	Identifying design requirements	By having the explanatory framework, this study can then identify the leverage points in the data analytics process where support can be provided to increase effectiveness or to reduce inefficiency of the users. As the result, a list of design requirement is formulated. This research activity aims to address ***Research Objective B***, which is to develop a conceptual framework for understanding the processes and requirements of actionable insight Corresponding content: Chapter 4
(4) Design Framework Development	Developing the design framework	The purpose of the design framework is to give a big picture of how the design effects work together to achieve the design goal. The framework consists of the theorized relationships between the designs and their design effects. And hence, it also provides the theoretical model for formulating the testable propositions. Three sub-activities are presented as the activities below. Corresponding content: Chapter 5
	Set detailed design objectives	With the design requirements, specific design objectives are formulated to address the design requirements. Each design objective specifies the design effect required to meet a design

	requirement. Together, the design objectives specify a coherent set of design effects needed to be delivered to meet the overall design goal. Corresponding content: Chapter 5
Developing design principles	This activity involves developing the design principles, which are the design guidelines for achieving the specific design objectives. They are the detailed blueprint for operationalizing the conceptual design into tangible system features. For each design principle, multiple alternative designs were proposed. This is important to prevent tunnel vision, solely focusing on single design from the beginning. The alternatives are evaluated in terms of this desirability in accordance to the design objective, technical feasibility, and anticipated outcomes. This research activity aims to address ***Research Objective C***, which is to create a design of data analytics system that supports the processes and requirements Corresponding content: Chapter 5
Hypothesizing design effects	With the concrete design supported by the design principles, this study hypothesizes the design effects induced by the design framework. Corresponding content: Chapter 5
Developing working prototype	This activity involves the development of the high-fidelity prototype. The prototype is a physical implementation of the design principles, for testing purposes. The prototype is also an expository device for conveying the design ideas, which is particular useful for practitioners. Details of the implementation are not included in this dissertation. The contents may be released upon being requested.

(5) Design Evaluation	Design the evaluation	This activity involves designing the evaluation for testing the testable propositions. In this study, a user study was conducted to collect data for the evaluation. This activity included recruiting the participants, designing the tasks for the user study, developing the measurement instruments, and conducting the pilot test. This research activity aims to address ***Research Objective D***, which to evaluate the proposed design of data analytics system Corresponding content: Chapter 6
	Evaluating the design	The purpose of this activity is to evaluate the working prototype in order to test the propositions about the design and the intended effects. Based on the result of the evaluation phase, the design principles are verified. Design principles that are proven will be translated into guidelines. As research tends to discover more detailed understanding and implications of the system after experimentation and interview, it is important to include these high-granularity findings into the design guidelines. This study contends that merely abstract design guidelines very often cause confusion and ambiguity, rather than give freedom of design. It is imperative for design guidelines to cover high-level guidelines and also high-granularity information, to ensure the intended result can be reproduced correctly. Corresponding content: Chapter 7, 8

3.5 Data Collection & Analysis Methods

Most of the design science research involves the validation of the design. The design can be in the form of theory, framework, system, or a set of functionalities. It is a common practice for design science research to evaluate the tangible system, which can be in form of low-fidelity prototype, high-fidelity prototype, or a complete system (Arazy et al., 2010; Carlsson, 2010). Such evaluation requires the researchers to collect primary data through various data collection methods that involve the use of the system by its intended users. The data collection and data analysis are determined by the research objective, that is, depends on what is the study trying to claim.

An objective of this study is to propose a new set of design principles which hypothesized will support the users to perform their problem-solving activities more effectively. As informed by theory and evidence from prior studies, users commonly undergo the problem-solving activities, even in absence of the supports (Endsley et al., 2011). Therefore, the objective of the evaluation is to observe the net effect that the design principles introduce. In such a context, the performance of the design principles can only be measured if there is a comparison between the system that is built on the design principles and a conventional data analytics system. This implies that 1) the evaluation is needed to be done at the system level and 2) a comparison assessment is needed to measure the net effect of the design principles.

The evaluation is needed to be done at the system level due to the abstract nature of the design principles. The design principles cannot be evaluated directly in its textual form because the subjectivity and the varied interpretation are likely to lead to different understanding by the users. More often, what the users perceived from the design principles are deviated from the design principles that is intended by the researchers. In contrast, the design principles are objectively represented when they are infused into a tangible system. It minimizes the variability of how the users perceive the design principles and thus allows the results across different users to be objectively compared and evaluated.

Chapter 4
Developing the
Conceptual Explanatory Framework

4.1 Overview

Despite actionable insight being widely recognized as the outcome of data analytics, there is a lack of a systematic and commonly-agreed definition of the term. More importantly, most of today's data analytics systems have failed to deliver *actionable insight*, creating the impression that actionable insight is just marketing hype, or another cliché word in the field of data analytics.

This study asserts that the root of the problem is that existing systems do not effectively support the problem-solving activities required to achieve actionable insight. Such ineffective designs are due to the lack of a systematic and theory-driven understanding of how users can achieve actionable insights. As implied, most data analytics systems to date are the result of advancement in data and computational techniques. These systems were designed with very little understanding of how the human users solve analytics problems.

In accordance with these gaps, this study contends there is a need for a theoretically driven and comprehensive understanding of actionable insight that can be used to 1) define actionable insight and 2) understand the processes and requirements needed to achieve actionable insight. In this chapter, Section 4.2 explains how this study conceptualizes the term "actionable insight" based on the findings from the literature review. Then, in Section 4.3, this study proposes a definition of actionable insight based on that conceptualization. Section 4.4 introduces a conceptual framework that is developed based on the way this study conceptual actionable insight. The framework helps this study to systematically understand the problem-solving activities required to achieve actionable insights. It draws inferences from the justificatory theories reviewed in Chapter 2: situation awareness, sensemaking, and the complex problem situation. Based on that framework, a set of design requirements is derived to effectively support the ways users solve complex analytics problems. The last section of this chapter summarizes the chapter and provides an overall discussion of the design requirements.

Figure 18 shows a simplified version of the research flow. Note that Section 4.2 and 4.3 addresses Research Objective A -- to propose a systematic and theory-driven definition of actionable insight. Section 4.4 addresses Research Objective B -- to develop a conceptual framework for understanding the processes and requirements of actionable insight.

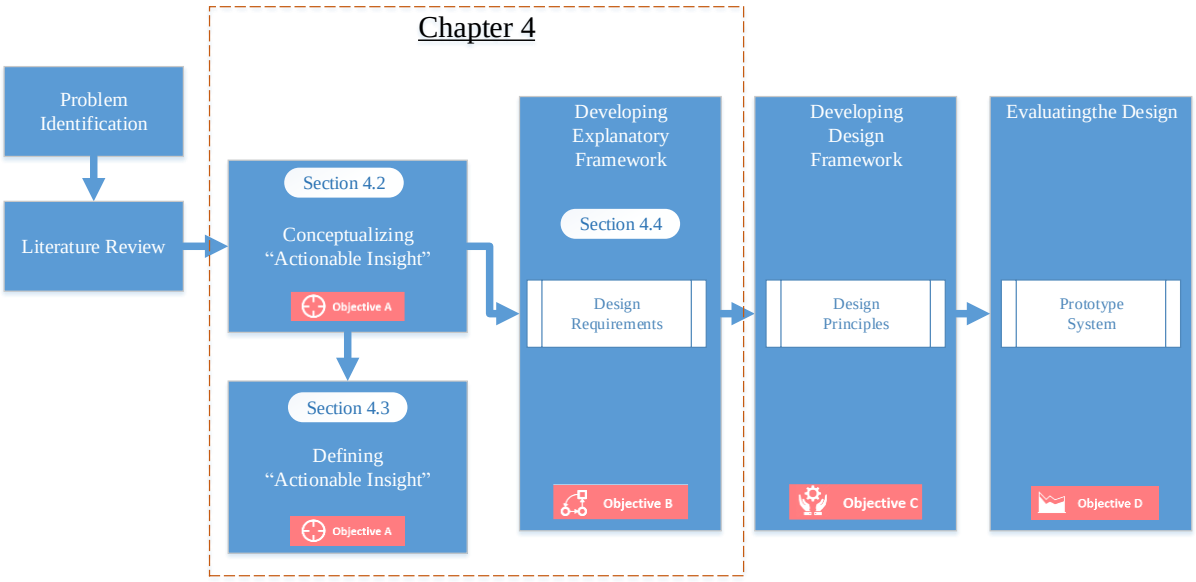


Figure 18 Contents of Chapter 4 and Research Objectives

4.2 Conceptualizing Actionable Insight and its Components

This study contends that, in order to systematically define actionable insight, it is important to first understand actionable insight as a concept. One common theme in the somewhat fragmented definitions of actionable insight describes it as the knowledge which enables users *to act upon meaningfully*. In the context of data analytics, this study defines the insight's capability to be acted upon meaningfully as enabling users to solve their analytics problems, on the basis of a coherent set of knowledge about the analytics problem.

Using this notion, this study conjectures that actionable insight comprises a collective of knowledge states about the analytics problem situation that enable the users to solve the problem. For the actionable insight to occur, its components - the knowledge states need to be achieved by the users. Through the review of situation awareness (SA) theory, this study asserts that the theory is suitable to be used to conceptualize the knowledge states of actionable insight. The reasons are 1) situation awareness is the outcomes of problem-solving activities that are directed toward actionable insight (Endsley & Jones, 2011), thus it is conceptually similar to the knowledge states in data analytics, 2) the theory can provide a theoretical basis for organizing the knowledge states and explains their relationships, 3) the theory can be combined with sensemaking theory to provide complementary explanations to holistically understand of the problem-solving activities required to achieve actionable insight, and 4) the theory is useful for informing designs. This is because it explains the user behaviors and cognitive states in problem-solving activities, and thus allows this study to identify the leverage points where user performance can be enhanced.

The review of the situation awareness theory revealed that the six awareness states from the theory can be used to explain the three major types of insight commonly found in analytic-related literature. The awareness states are 1) identification awareness, 2) perceive awareness, 3) integrative awareness, 4) comprehensive awareness, 5) foresight awareness, and 6) simulative awareness. *Figure 19* recapitulates how the awareness states relate to the types of insight found in analytic-related literature.

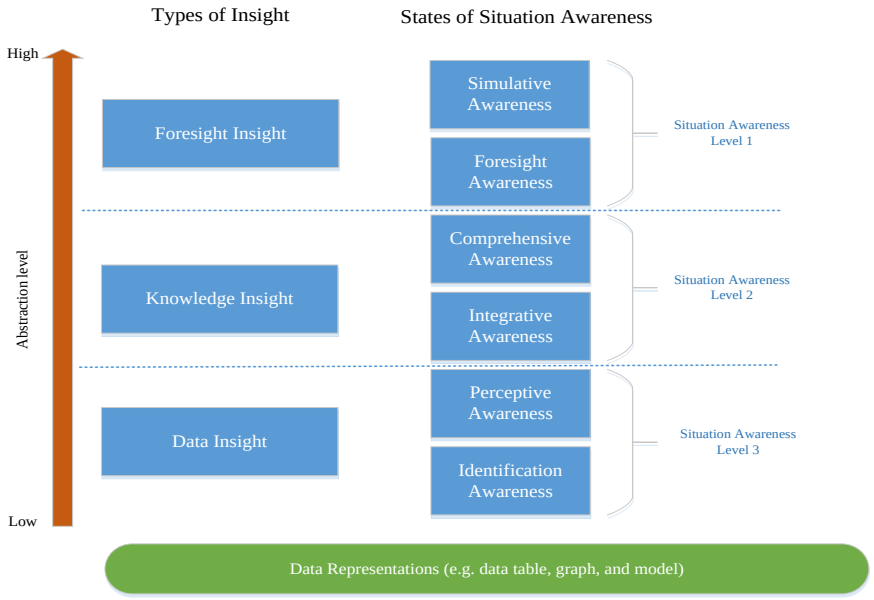


Figure 19. Major types of insight and the states of situation awareness

Using the structure of the six-state situation awareness, this study breaks down the three major types of insight into six states of insights, namely 1) identification insight, 2) perceptive insight, 3) integrative insight, 4) comprehensive insight, 5) predictive insight, 6) prescriptive insight. These six insights are the knowledge states that collectively constitute and define actionable insight. *Figure 20* shows the conceptual structure of actionable insight. It consists of the six insight components, each belongs to one of the three major insight types. These three major insights, analytic insight, synergic insight, and prognostic insight, correspond to the main phases of the data analytics process 1) data exploration, 2) information synthesis, and 3) knowledge actualization. Overall, such decomposition allows the insights to be examined in the way that is well aligned with the process and outcome perspectives from the justificatory theories. The structure enables complementary explanations for understanding actionable insight to be drawn from the theories.

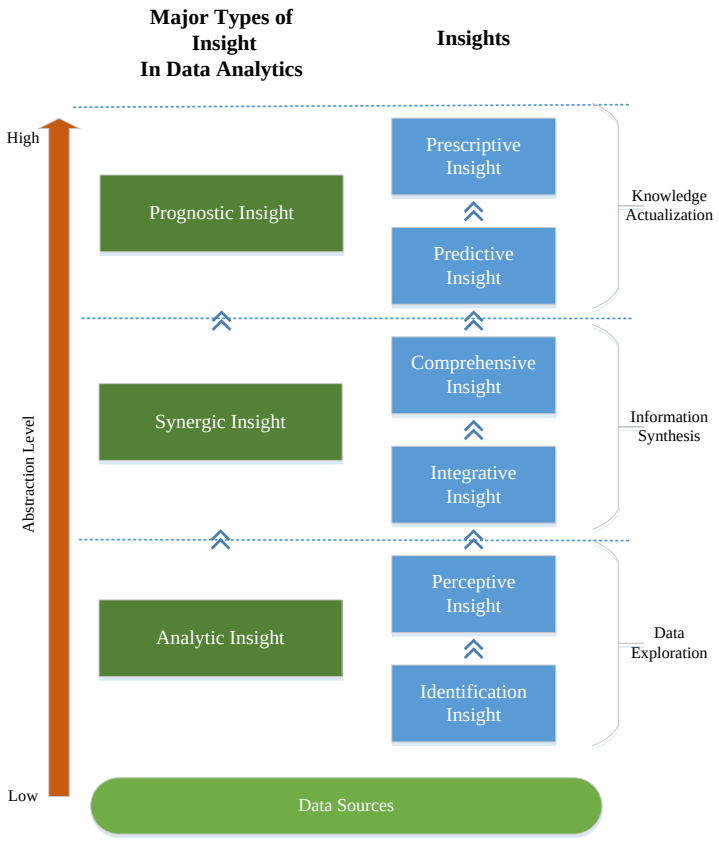


Figure 20. Insights in Data Analytics

Note that this study uses the term "insight" rather than "awareness", to differentiate these components of actionable insight from the constructs in Endsley's situation awareness theory (Endsley, 1995b). These insight components are specifically developed for the field of data analytics. As opposed to the six-state situation awareness which focuses on a user's cognitive awareness states in general tasks, the six derivatives specifically describe the information-processing states that have resulted from the human-information discourse in data analytics. As the insights in data analytics, these states partially exist in the physical world and are partially maintained within the user's cognition, as opposed to situation awareness states that are entirely in the user's cognition. This form of distributed cognition is a more realistic representation of the data analytic process as it implies that users do not maintain, process, and store all the knowledge in their mind (Zhicheng, Nersessian, & Stasko, 2008). The knowledge can be processed and applied for problem-solving on the fly when the users interact with the external counterparts of these knowledge states.

4.3 Defining Actionable Insight

In this study, actionable insight is conceptualized as a multi-component concept which consists of three major components, namely analytic insight, synergic insight, and prognostic insight, as shown as the following:

- **Analytic insight** – understanding and interpretation of individual analytical results.
- **Synergic insight** – comprehension of the connections between the analytic insights and understanding of the problem situation as a whole.
- **Prognostic insight** – prediction of the problem situation's future states and the assessment of their effects on the problem situation, objectives, and constraints.

Considering on the three major components of actionable insight, this study proposes a formal definition of actionable insight as the following:

> **Actionable Insight:** *A set of progressive knowledge about the analytics problem situation, based on prognostic insights derived from synergic understanding of analytical results, which enable the user to make an informed decision to solve the analytics problem.*

Based on this definition, actionable insight is therefore the coherent states of knowledge the users have gained at different stages of the data analytics process. In essence, actionable insight is the understanding of the analytics problem which enable the users to answer 1) what is happening, 2) why it is happening, and 3) what will happen next. Together, these progressive understandings provide the user with the sufficient understanding of the problem situation in order to decide on a solution that is best suits the user's objectives and anticipated scenarios. This thereby constitutes the "actionable" notion of the term. The definition of actionable insight in this study also suggests that it is:

- ***A reasoning artifact***. That is, a product of a user's analytical reasoning process, which based on the outputs from a series of analyses. During the process, the technical outputs could be interpreted, synthesized with each other or with existing knowledge, schematized, simulated mentally, and scripted with certain action plans. Hence, actionable insight is not the immediate results from the analytics system, but the high-level knowledge derived through the interaction, internalization, and reasoning between the analytics result and the users.
- ***Practical knowledge.*** That is, a set of cohesive knowledge that can be operationalized in such a way as to guide decision or action towards an intended objective. This implies

actionable insight entails pragmatic implications or value. It is a context-specific knowledge which can be readily applied to affect the user's job or decision. Its notion of actionable largely depends on whether a user can deploy the knowledge learnt from the data analytics to solve a practical problem.

- ***Comprehensiveness of understanding***. Based on the notion of actionable insight in this study, actionable insight is not a state of either exist or non-exist, but actionable insight is the extent of comprehensiveness to which the data analyst's understandings of the problem situation that allow him or her to decide on an action plan that can be used to solve the problem situation. This connotation implies that actionable insight can be measured based on the comprehensiveness of the insight components. This indicates that the more comprehensive the insight components are, the greater likelihood there is that overall understandings of the problem situation will allow the users to make informed decision.

4.4 The Hierarchical Framework of Insights

For the definition of *actionable insight* to be creditable and theoretically sound, the definition must be built on a full-blown concept that provides thorough understanding of the phenomena being defined. Therefore, a conceptual explanatory framework is developed as the theoretical foundation for the term actionable insight. The conceptual explanatory framework, named *Hierarchical Insights in Versatile Environment* or HIVE, describes the composition of actionable insight. HIVE suggests that the insight components are hierarchically related. Components at the upper layer are built upon the components at the lower layer. *Figure 21* illustrates the simplified view of the HIVE framework with the positions of the major insight components.

---- This space is intentionally left blank ----

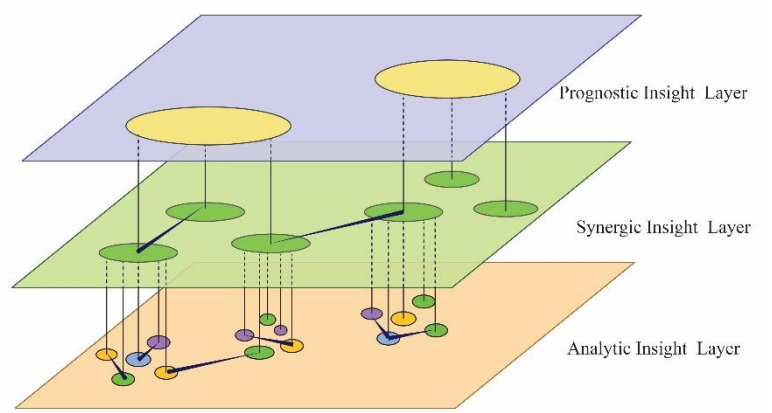


Figure 21. Simplified representation of the conceptual explanatory framework

The upcoming subsections describe each of these major insight components in detail, starting with analytic insight and ending with prognostic insight. Each subsection contains discussion about two of the six insight components. Each of the discussion first explains the insight component and involved activities, then identifies the design requirement(s) based on aspects which can be improved.

4.4.1 Major Component 1: Analytic insight

Analytic insight is achieved when users successfully identify and interpret a relevant aspect of the data during the data exploration phase. Analytic insight is what practitioners and researchers commonly refer to as "insight", which is an observation about the data. Examples of analytic insight could be the understanding of key information from the enquiry results, such as a set of association rules, a significant pattern in a visual graph, or a relationship within a multi-regression model. Achieving analytic insights means that the users have transformed the data into useful information. At this stage, however, the individual pieces of information are not organized, grouped, or related to each other. Often, there are quite a number of analytics insights gained by the users at the end of the data exploration phase.

Each analytic insight is a relevant observation derived directly from one or more enquiries. *Figure 18* shows the relation between analytic insights and enquiries. An enquiry can exist in the form of a database query, a visualization, or a mathematical computation that transforms data into data representations. An analytic insight may also be derived from a series of sequential enquiries, in which the latter enquiry is built on the former counterparts. For instance, data is first being clustered, and then the resultant clusters are used as inputs for association rule mining. Analytic insight can be easily quantified and traced back to the enquiries in which it is being observed (Saraiya, North, & Duca, 2005).

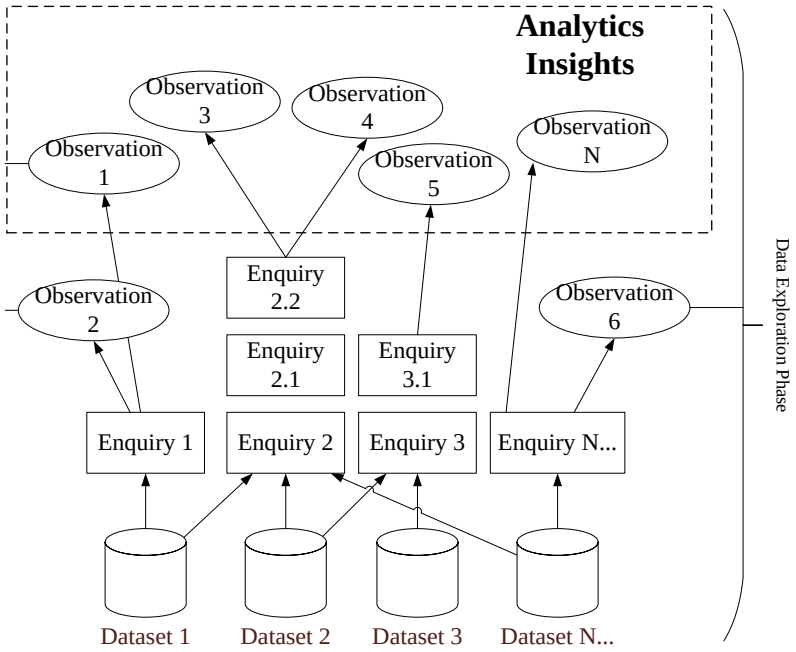


Figure 22. From dataset to analytic insight: identification + perceptive insights

Among the three major insight components, analytic insight is a low-level insight. It involves highly objective information that requires little judgmental heuristic. Each analytic insight has a narrow scope, as it focuses on a very specific piece of finding in the entire problem situation. Analytic insight reassembles the notion of using a high-power telescope to zoom at a part of a city. It allows you to see extreme details of a building, but it is difficult to get an overview of the city layout and to find your way to the building. Alone, analytic insights are unlikely to carry direct implications for decision making or allowing meaningful action. In other words, the practical value of analytic insight is generally low. Most of the data analytics systems such as data mining, data visualization, and statistical analysis cease their supports beyond this point.

As aforementioned, analytic insight is achieved when users successfully identify and interpret a relevant aspect of the data. This implies that the *relevant aspect of the data* needs to be identified before the *interpretation* can occur. In this notion, analytic insight comprises of 1) identification insight, and 2) perceptive insight.

4.4.1.1 Identification Insight

Identification insight is gained when users have successfully identified relevant data elements (Endsley & Jones, 2011; Lu et al., 2012). During the data exploration phase, users need to search and filter relevant data elements from large numbers of available data elements. For instance, the data available for stock market analysts could range from micro-level data such as company financial statements, internal reports, and company news, to macro-level data such as industry trends and macro-economic indicators of various countries. The outcomes of the search & filter activity are the relevant data

elements which are then used to generate the enquiries, as shown in *Figure 23*. The selection of relevant and appropriate enquiry techniques is beyond the scope of this study. That topic has been widely studied by researchers in relevant areas such as data visualization, data mining, and statistically analysis.

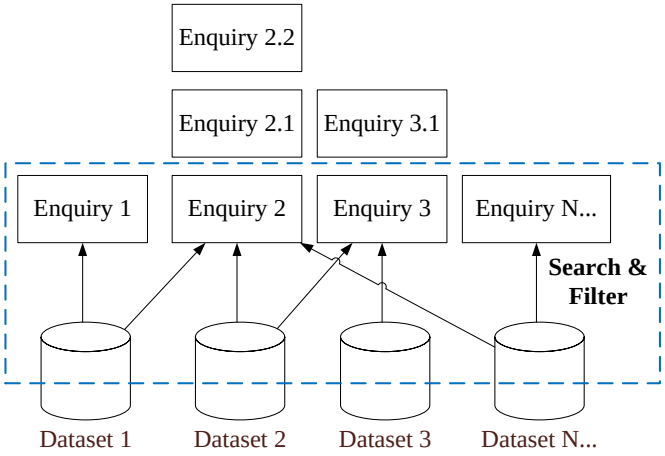


Figure 23. Identification insight involves search & filter activity

General users are very inefficient at dealing with large and complex data due to the limited cognitive capability. This challenge is particularly manifested in complex analytics problems. At the early stage of the analysis, users often do not know what data to explore and everything could seem relevant **t**o the analysis. Under such conditions, users commonly explore data based on hunches and experiment with the data through time-consuming trial-and-error processes (Heer et al., 2008; Heer & Shneiderman, 2012b). In such an opportunistic approach, users consider only the information discovered by chance, and resulting in mediocre solutions to the analytics problem. Research has also shown that, when overwhelmed by too many selections, users tend to reduce the amount of environmental scanning (Chen & Lee, 2003). They try to deal with the complexity by going to the same sources of information that they feel comfortable with or with which they have had positive results in the past (Weick, 1995).

Consequently, they tend to rely on fewer data sources and prematurely narrow down the data exploration. As a result, they are likely to simplify assessment of the problem situation, which eventually leads to poor problem solutions. The problem with the simplified situation assessment approach is particularly dangerous in the complex problem situation for several reasons. Firstly, each complex problem in practice is arguably unique and highly unpredictable (Mirel, 2004). These characteristics tend to render previous assumptions becomes invalid. The second reason is that the incompleteness and inaccuracy which resulted in the early phase of data analytics may be propagated and magnified throughout the rest of the process. Research has found that exclusions of data elements based on prior experience or intuition of the analysts has often led to misleading solutions. These solutions may appear to be highly sensible but have often failed miserably to address the problem. Worse, they overinflate the analyst's confidence in the solutions. Therefore, there is a need to support users to effectively explore the data to identify relevant data elements.

Design requirement: To support users to effectively explore large numbers of data elements.

4.4.1.2 Perceptive Insight

Enquiries do not automatically transform the data into information. They require the users to interact with the data since only the users can determine the context, meaning, and relevancy of the key information from the enquiries. This interaction involves 1) receiving the information through the visual perceptual sensory system and 2) interpreting the data into a mental understanding of a phenomenon which the information conveys (Meyer, Thomas, Diehl, Fisher, & Keim, 2010). Therefore, perceptive insight involves a perceive & interpret activity that transforms the data into meaningful information.

Perceptive insight is gained when the users have successfully perceived and interpreted an aspect of the problem situation. In other words, each perceptive insight is a *meaningful* and *relevant* observation made from one or more enquiries. Note that only observations that are both *relevant* and *meaningful* to the current analytical task are considered as perceptive insights. Therefore, it is possible to discover many meaningful observations that are "good to know" but only selective ones of these are relevant to the currently analytical tasks. Each observation is a semantically meaningful statement that describes a specific aspect that is relevant to the problem situation, derived from facts learnt from one or more data elements. For example, observation can be "stock price of company B has plunged 30% in the past quarter". *Figure 24* shows the relationships between enquiries and the observations. A single enquiry can result in more than one observation (e.g. observation 2 and 3). Likewise, an observation can be made from one or more enquiries (e.g. observation).

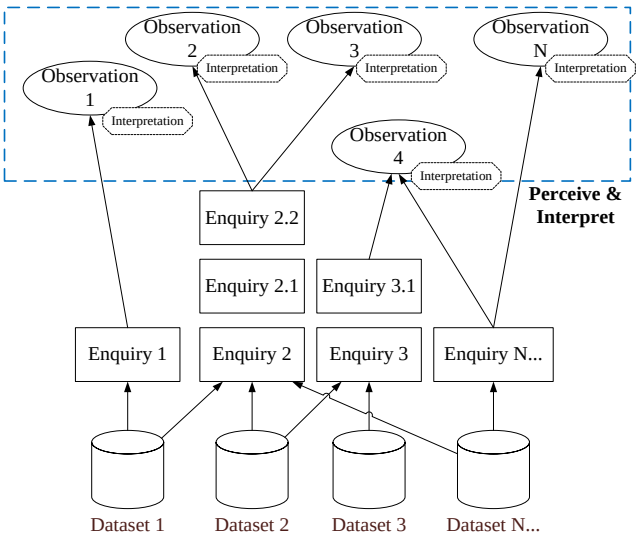


Figure 24. Perceptive insight involves perceive & interpret activity

What is critical in an observation is not just the meaningful statement about a specific aspect that relevant to the problem situation, but also the interpretation of how the statement is derived. (Lefebvre,

2004) also pointed out that in complex problem solving, it is important to go beyond dry data analysis. It often requires the interpretation of the data. Conceptually, the interpretation can comprise of two main types of information 1) facts learnt from the enquiry of data elements and 2) the rationales used to justify the observation statement. The facts are the key attributes for supporting the observation statement. The facts learnt can be further categorized into: 1) data attribute and 2) derived attribute. A data attribute is based on the value that is explicitly reflected by the enquiry result, such as 30% sales increase or a quarter starting 1st July. A derived attribute is the value derived or inferred based on the user's subjective judgement, such as a *strong* or *weak* increasing trend. Multiple attributes may be combined to provide a greater context for the observation. On the other hand, the rationales describe the reasoning, flow of logic, and context that the users used to extract the information from the enquiry. For example, the 30% sales increase is a strong trend because the sales increments of the company over the last 5 years were between 5 to 10%. *Figure 25* illustrates the visual representation of the structure of an observation. Note that an observation does not necessarily consist of all these components. At the bare minimum, an observation should contain at least the observation statement. This is because, in practice, users may have difficulty to clearly explaining how they derive an observation.

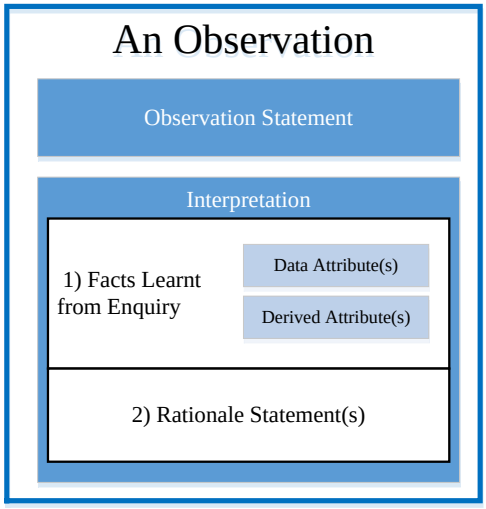


Figure 25. Conceptual structure of an observation

A common criticism from academic research on existing data analytics systems is the lack of their capability to save and manage observations (Ling, Gerth, & Hanrahan, 2006; W. Wright et al., 2006). This shortfall posts a great challenge for users who are dealing with complex analytics problem. *Figure 26* shows the three main reasons why observation management is important in such a situation. Conceptually, each blue dot represents an observation discovered. Firstly, the analysts often make a large number of observations during the data exploration phase. Secondly, data exploration is a messy and opportunistic process where the analysts constantly switch between different information perspectives to discovery meaningful observations (Mirel, 2004). Thirdly, added to the complication is the highly dynamic nature of the data exploration process. The resulted observations are unstable

because 1) they could be just the temporal stepping stones to reach the next observation, 2) they need to be updated later, or 3) they are could be found irrelevant in the later stage.

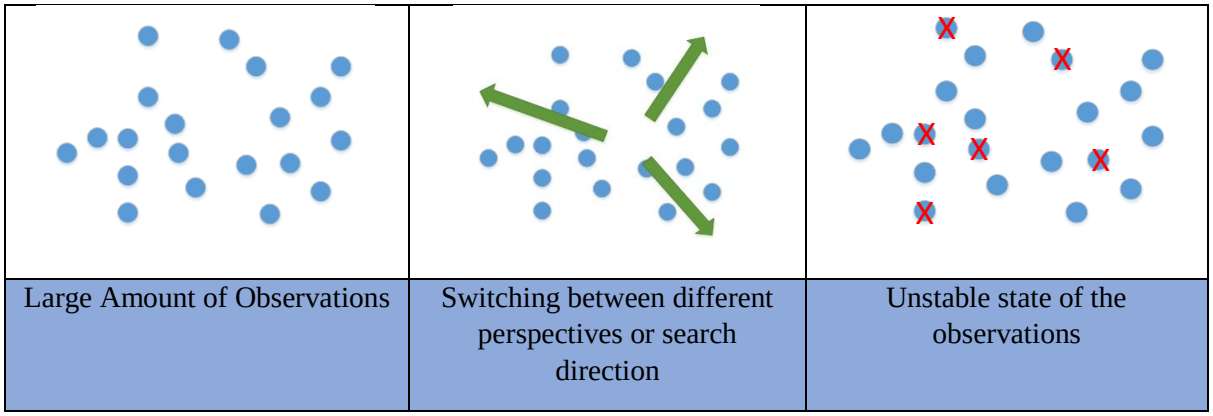


Figure 26. Characteristics of observations in complex problem situation

Given these three characteristics of the observations, the cognitive efforts required to accurately recall the observations tend to exceed the cognitive capacity of the users. Researchers have pointed out that the cognitive capacity used to keep track and recall the observations is exhausting the same pool of capacity for reasoning, thereby impairing the reasoning performance. Jones and Endsley (1996) found that working memory losses – a phenomenon in which the observations was initially perceived and then forgotten - is one common error that can jeopardize the overall analytical performance.

This is not to mention the interpretation underpinning each of the observations. Without support, the users often derive observations on the fly and forget about the interpretation as soon as they move on to the next observation. However, the interpretation is important as it exposes the assumptions and interpretations of an observation. It allows the traceability of the observations, and thus accounting for the creditability of the analysis. Most existing systems do not support users in capturing the assumptions and interpretations of the observations**.**

Without support from the systems, users often have difficulty in correctly recalling the observations. They might have to rerun the enquiries just to recall what they previously found. A significant amount of time and efforts is often wasted during the data exploration. They might fail to rediscover the observations if they have forgotten the interpretation they previously used. Moreover, the communication between two analysts about their observations is not easy because the interpretations are often implicit. One analyst may find it difficult to understand how the other analyst derive a particular observation. This weakness is signaling that there is a need to support the users, to generate meticulous, traceable, and defensible interpretations.

Design requirement: To support users to capture, manage, and retrieve their observations, including the underlying interpretation.

Researchers have also been criticized existing data analytic systems for their lack of support to allow the extended use of the observations (Thomas & Cook, 2005). Data exploration commonly results in a great number of observations about diverse aspects of the problem situation. Next, the data analysts often need to derive a joint summary from these observations and proceed to higher-level analytical activities. For instance, analysts could derive separate lists of desirable stocks-based analysis of different data elements such as financial health, market reputation, director board, and new product plans. It is important for them to be able to have a summarized view of their observations in order to identify a few potential stock options to be examined in-depth. Yang et al. (2009) pointed out that it is common that the data analysts go through the observation one by one in attempt to extract important findings across a number of observations.

This practice is a highly ineffective yet counterproductive process, given that there will be a sharp drop in the human reasoning capability when the observations that they need are not accessible by them (Yang, Jing, & Ribarsky, 2009). When only one observation can be shown at a time, the users have to mentally retain the information of multiple observations and attempt to derive a joint conclusion based on the slowly fading information. There is a limit to how many observations they can consider at a time, and thus makes the process difficult to scale up to a complex problem situation. Ensley and her colleagues (2011) have also found that one common cause of the situation assessment errors is attributed to the non-presence of the information at the point of time when users need it for reasoning. As a consequence, the accuracy of the joint conclusion is likely to be jeopardized and could significantly deteriorate the quality of the data analysis, as the errors are compounded through the data analytics process.

Design Requirement: To support users to create joint summaries from their observations.

4.4.2 Major Component 2: Synergic Insight

Synergic insight is the comprehension of the problem situation as a whole. It happens during the information synthesis phase of data analytics, which has been regarded as a transitional stage between data analysis and actionable insights (Robinson, 2009). During this stage, low-level information is integrated and synthesized with human knowledge to form high-level knowledge about a significant aspect of the problem situation. Then, the high-level knowledge is connected to build a knowledge network that facilitates the understanding of the problem situation as a whole. The synergic insight is larger than the sum of its individual components because it allows users to understand the dynamicity

of the situation in a single big picture. Studies have stressed that the process of creating the big picture is a prerequisite for solving complex problems (Pohl et al., 2012).

Figure 27 illustrates the conceptual structure of a synergic insight. Note that the relation between analytic insights and a synergic insight is intermediated by the chain of arguments (surrounded by the dotted rectangle in the middle tier). Viewing from bottom to top, the figure shows that the analytic insights discovered in the previous data exploration stage are integrated to form a chain of argument. Subsequently, multiple chains of argument are connected to form the complete big picture of the problem situation.

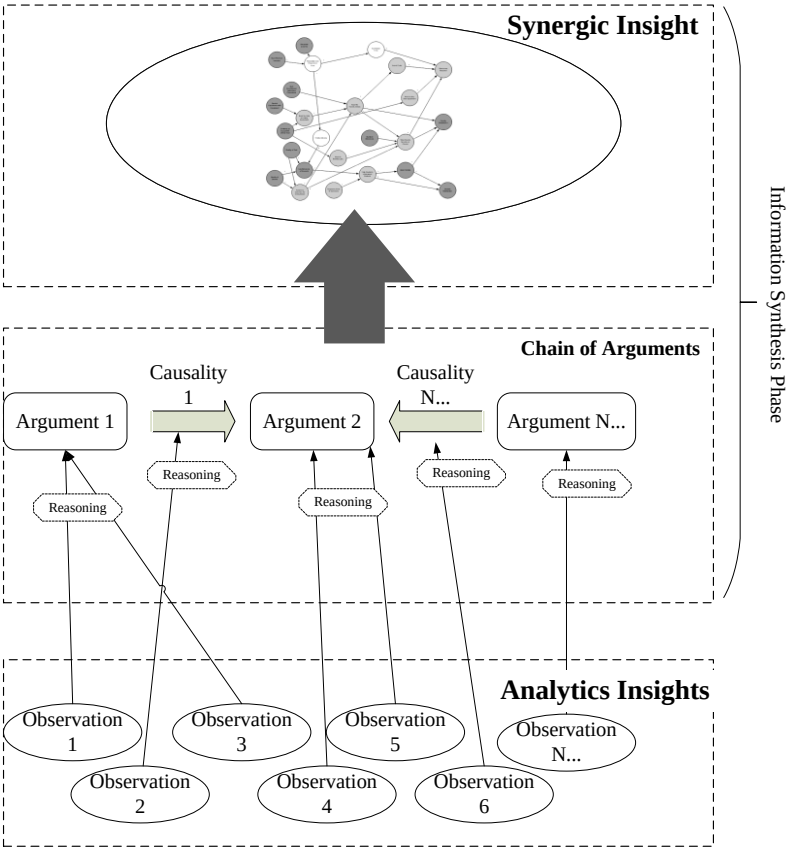


Figure 27. From analytic insights to synergic insight: integrative + comprehensive insights

Synergic insight does not readily warrant a decision. At this stage, users understand the problem situation as a whole based on the collective interactions among the key entities or factors within the problem. However, they do not know what the possible outcomes of the situation model would react to different solution alternatives or scenarios. Synergic insight is the vital foundation for users to design their solutions, and to prepare for the most of important stage of problem solving discussed in the next section.

Synergic insight is achieved when users 1) create high-level knowledge that is key for understanding the problem and 2) comprehend the problem situation as a whole. This implies that there are two components of synergic insight, namely integrative insight and comprehensive insight.

4.4.2.1 Integrative Insight

Solving an complex analytics problem requires a high-level conceptual understanding of the problem that goes beyond dry data analysis (Garg, Nam, Ramakrishnan, & Mueller, 2008). When analyzing and solving a real-world problem, users commonly think, define the problem, and seek for a solution at the concept-level, as opposed to the data-level (Lefebvre, 2004). The analytic insights are often fragmented and the relationships are obscure. In order to make use of the analytic insights they have found, users need to understand the relationships among the pieces, and integrate them to create higher-level knowledge (Yang et al., 2009). An empirical study conducted by D. Gotz and Zhou (2008) has reaffirmed that more than three-fourths of the users develop higher-level knowledge from the low-level analysis findings.

Integrative insight is gained when users have successfully integrated the separated analytic insights and synthesized them with their subjective knowledge to form a high-level knowledge that is meaningful at the problem-solving level. Conceptually, integration happens when more than one analytic insights are combined to form a higher-level conclusion. On the other hand, synthesis happens when one or more analytical insights are combined with implicit and subjective knowledge that is not available in the system. However, this study asserts that it is difficult to clearly distinguish integration and synthesis in practice because they often happen together. The integration and synthesis activities result in the integrative insight that moves beyond information recall or adding up information. Integrative insight is a high-level knowledge about a specific actor, entity, or factor within the problem landscape. Figure 28 shows how the knowledge is derived from observations.

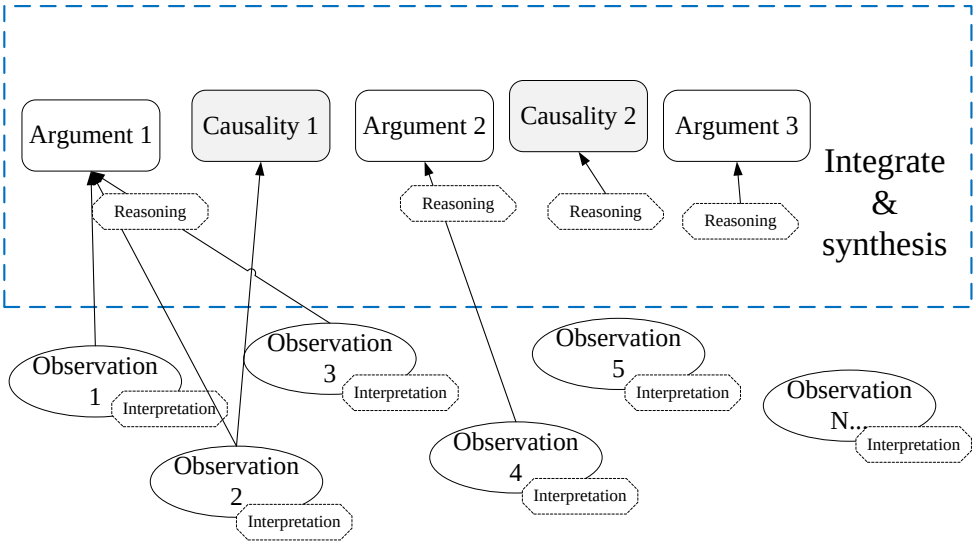


Figure 28. Integrative insight involves integration and synthesis

In complex analytical tasks, the high-level knowledge often exits in the form of "argument". This is because the knowledge is often a logical inference that combines facts and the users' subjective

reasoning into a defensible judgment of greater knowledge value (Thomas & Cook, 2005). *Arguments* and *association* in *Figure 28* are the high-level knowledge.

Table 7. Argument types

Argument	Logical inferences combining fact-driven *observations* and subjective *reasoning* into defensible judgment of greater knowledge value.
Association	A specialized type of argument that make inference that an argument (e.g. actor, entity, event, or factor) is associated with a second argument.

Every argument is a defensible statement about a specific factor in the problem situation. It draws on one or more relevant analytic insights as the supporting evidence. An example of argument can be "the decrease of competitive advantage in accounting software market". This argument can be supported by several analytic insights such as a) a consistent drop in market shares, b) the significant increase of unsatisfied customers, and c) a sharp increase in customer churn rate. Argument 1 in *Figure 28* shows the structure of such an example. Association is a specific type of argument that make an inference about one argument that is associated another argument. For example, when inflation rate is increase, the export rate of electronic industry is also expected to increase.

Besides the data-driven analytic insights, subjective reasoning can be synthesized with the insights to form an argument. There are three ways the synthesis can happen: 1) directly supporting an argument, 2) justifying how multiple analytic insights are integrated to form an argument, 3) describing the key attributes of an argument.

For the first, reasoning can be included to jointly support an argument. Reasoning can be the users' knowledge, experience, judgement, assumptions, or other external information that is not stored in the system. For instance, reasoning can be the consensus that "our existing software will fail to meet the emerging industrial practice in the next 2 years", resulting from a meeting among the managers. Note that it is possible that an argument is formed merely based on reasoning (which illustrated by Argument 2 in *Figure 28*). This is especially true in real-world problem solving, where certain arguments are vital, but yet the data to support them are impossible to obtain.

Secondly, in addition to directly supporting an argument, it is also common that the reasoning acts as the glue that integrates the individual analytic insights into a cohesive argument (David & Michelle, 2009). In such a case, the reasoning is the justification of how the user infers the argument from several analytic insights. Lastly, the reasoning can also describe the key attributes of an argument. For instance, it can describe the confidence level that the users have towards that argument. *Figure 29* shows the conceptual structure of an argument or association which includes three types of reasoning.

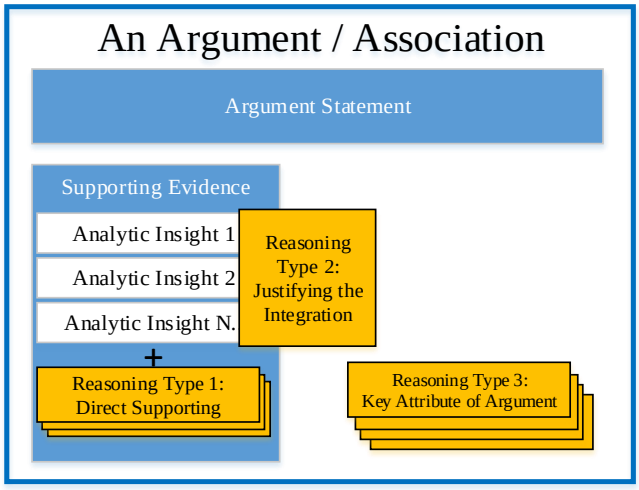


Figure 29. Conceptual structure of an argument / association

Researchers have commented that despite the widely recognized importance of knowledge creation, most data analytics systems do not explicitly support the integration and synthesis activities required to create the knowledge (William A. Pike, Stasko, Chang, & O'Connell, 2009). Most existing data analytics system explicitly support users only up to discovering analytic insight, the integration and synthesis are mostly manually done, either in the users' mind or with paper and pen.

The manual integration and synthesis processes are particularly taxing on the analyst's cognition. This is because constant and dedicated efforts are required for 1) recalling and reviewing relevant analytic insight from all the observations made, 2) integrating the analytic insights and synthesizing them with users' domain knowledge, and 3) keeping track and managing the newly created knowledge. These three processes are competing for the same pool of cognitive resource, and often exceed the capacity of user's cognitive resources. As a result, this impedes the reasoning capability of the analysts, prompting to errors and biases in the derived knowledge (Yedendra B. Shrinivasan & Wijk, 2008). Despite the quality of the individual analytic insights, they would be worthless if the users were not able to put them together in a meaningful way to aid problem solving. Therefore, the following design requirement is formulated.

> **Design Requirement:** To support the users to create, manage, and retrieve high-level knowledge that is derived from low-level analytic insights and reasoning.

4.4.2.2 Comprehensive Insight

As noted, each argument focuses on a single key factor in the problem situation, such as an entity, an actor, an event, or a concept. Individually, it provides only a fragmentary description for the entire problem situation. Complex problem solving is synergistic and not additive in nature (Mirel, 2004). For the human problem solver to understand the complete picture of the problem landscape, they need to 1)

understand how the entities, actors, events, or concepts are connected and 2) build a simplified replica of the problem, or a *situation model* (Greiff, 2012; Nakatsu, 2010). Therefore, comprehensive insight involves a connect & build activity. The construction of this situation model is the key to effectively solving a complex problem (Albers, 1999; Gary & Wood, 2011; Jonassen, 2000). The importance of such a big picture can been understood, as researchers have often claimed that when the structure of a problem is made clear, a solution can be found fairly easily (Pohl et al., 2012). This maybe is because an accurate situation model is the enabler of hypothesis generation, solution seeking, and other high-level cognitive activities that lead to effective problem solving (Ribarsky et al., 2009).

Comprehensive insight is achieved when the users have successfully built a situation model and comprehend the problem situation as a whole. *Figure 30* shows how the situation model is created based on the arguments and causalities. Firstly, it involves connecting the arguments through causalities. As a result, a chain of argument is formed. Note that the causalities now form the links between the arguments. An example of the chain of argument is a "decrease in competitive advantage in personal accounting software market (an argument)" will negatively affect the "reputation of professional account software market (an argument)" through "reducing reputation of related products" (an association).

It is common to have multiple chains of argument. Each chain of argument may focus on the logical chain of relationships within a particular area within the problem situation. For example, a chain of argument may focus on the marketing area, while another may focus on the production area during an evaluation of a company value in the stock market. Subsequently, multiple chains of argument are structured into a situation model to help users comprehend the entire problem situation. For instance, all the chains of argument involving different areas such as marketing, production, and finance, are joined in order to have collective effects on an objective variable of the users, say "value of the company". The overall structure of a situation model is largely shaped by the users' domain knowledge of how things work and related.

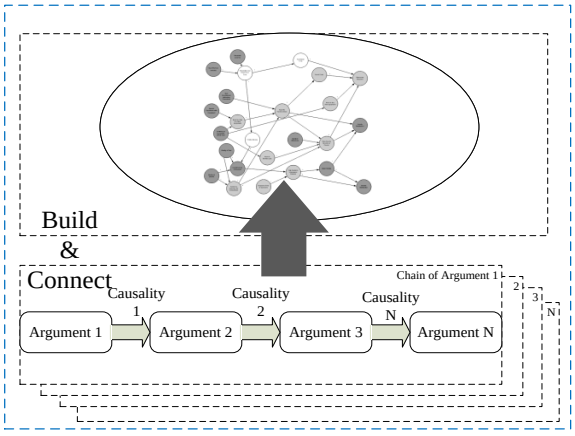


Figure 30. Comprehensive insight involving connect & build activity

The construction of a situation model is a highly challenging task. It is commonly known that human's cognitive resource is not capable of holding all the information and building a situation model without external aids. Yet, the ability to hold, view, and manipulate the factors within a working model is vital in a complex problem-solving process (William A. Pike et al., 2009). Researchers have even claimed that data visualization is the wrong primary tool where the formation of explanatory or correlative models is the desired outcome, and have asserted a need for "model visualization" rather than "data visualization" (Amar & Stasko, 2004). To the best knowledge of this study, there is no commercial data analytics system were found that support situation modeling, and only a few research works involved users to synthesizing information to create a situation model. See Section 2.2. The reviews have identified two significant limitations in those systems to support the users in creating the big picture of the problem situation. These limitations are:

- Difficulties for users to identify the structure of the situation model
- Over-reliance on manual and subjective reasoning to create the situation model
- Difficulties in creating a situation model that can go "live" – a dynamic situation model

A common question that users have in developing a situation model is "where do I start?". This hurdle faced by analysts is to identify the preliminary structure of the situation model. The structure of situation model contains two main components: 1) the individual constructs and 2) the relationships between the constructs. Studies have found that experienced domain experts have richer mental schemas from which they can draw inferences to create a core structure of the situation model. This core structure acts as a preliminary logical framework that helps the domain experts to integrate various constructs. As the outcome, they are able to construct a more complete and accurate situation model in lesser time. However, not everyone has the knowledge or experience to generate the preliminary core structure of the situation model. Most users would spend a significant amount of time deciding on the preliminary structure by repetitively building and scrapping structures until they have a satisfactory one. Moreover, Endsley (1995b) found that less experienced domain users may fall far short of being able to integrate various constructs in order to comprehend the situation, even when they have the same level of analytical insight.

> **Design Requirements:** To support the identification of the preliminary core structure of the situation model

Due to uncertain and incomplete information in complex problems, data analysts often need to integrate quantitative and qualitative information in a complementary manner to construct a complete situation model. In a complex analytic task, it is rare that all the information needed for the data analytic can be obtained from existing datasets. Much other key information, such as know-how, domain rules,

external information, opinions, and assumptions, exist in the form of qualitative information (Eppler & Platts, 2009). This qualitative information is often stored in the mind as implicit knowledge and is not available in the datasets. Nevertheless, qualitative information is critical input for situation modeling (Jonassen, 2000): it leads to the completeness, realisticness, and depth of the situation model. Quantitative information can be useful for statistically establishing and testing the relationships in a situation model.

Nevertheless, most of the systems that support situation modeling rely almost entirely on the analysts' subjective intuition and judgment to establish the arguments and their relationships. For instance, the scalable reasoning system (SRS) relies on the users to enter the confidence value of an argument and the strength of an association, based on their subjective understanding of the information. Although this allows the model to be easily specified using user inputs, a costly downside is that the quality of the resultant situation model relies largely on the assumption that these user inputs are accurate (Zuk & Carpendale, 2007). Nevertheless, every complex problem situation is often unique and novel, to a certain extent, so even experienced domain experts may not able to able to provide accurate inputs to the situation model. Moreover, users who form erroneous beliefs about the relationships between argument tend to make decisions on the basis of beliefs, even though numerous evidence indicate that the beliefs were wrong (Lee & Chen, 1997). Consequently, the quality of the situation model suffers. Another problem with using such an approach is that it forfeits the power of quantitative information, making the data collected only useful for indirectly informing the structure of the situation model, but not directly as the building blocks of the situation model. Moreover, it is almost impossible to assess the validity of the model created. The users would easily base their decision on an inaccurate situation model. Therefore, this study asserts that there is a need to support the creation of the situation model with a method that can take advantage of flexibility of subjective reasoning and the rigor of quantitative information.

> **Design Requirement:** To support both quantitative and qualitative approaches to situation modeling

Additionally, the review of relevant works also revealed that most of the systems are no more than a canvas that requires users to manually write down, organize, and connect the individual analytic insights. The features in these systems are meant to "hold" the information for the users, so the users can process the information. In other words, the features focus on alleviating the users from the "attention span" and "working memory" constraints, but do not enhance the reasoning performance of the users. The major drawback of the method is leaving human users to do the entire reasoning, which can be inefficient and subject to cognitive biases. For effectively facilitate the analyst's reasoning, the situation model has to be dynamic enough to enable rich interactions between the model and the analyst. The situation model must be able to be dynamically updated as the analysts progressively find new

information. More importantly, a dynamic situation model should take advantage of computer-aided reasoning techniques to enhance the reasoning of the users.

The challenge is that dynamic situation models require underlying data, logic, computation, and interaction mechanisms to respond to the user interactions and to use computer-aided reasoning techniques. As a result, this type of modeling often requires data analysts to go through a tedious and rigid approach to constructing the model. The analysts may need to learn to use specific modeling syntax to specify the model's structure and to be able to produce mathematical equations to represent the interactions between the components in the model. For instance, Bayesian Network modeling requires the analysts to be savvy in terms of statistics, mathematical modeling, and programming. This approach may reduce the productivity of data analysts because it requires them to spend significant time on the mathematical modeling, rather than on the actual data analysis.

Moreover, most users and analysts are often non-technical personnel, which prevents them from creating a firsthand situation model. Situation modeling is commonly being done is through interviews that involve domain experts and researchers. Then, the industrial or academic researchers translate the domain experts' requirements into a situation model. The major downside of this method is that the process takes a long time and significant effort. Researchers have reported that such situation modeling tasks can potentially last for weeks or even months. The heavy costs often lead to one-off situation modeling projects where its structure is unlikely to be updated and the model is often limited to illustrating a snapshot of the problem situation at that particular point of time. Such situation models are mostly for research reporting purpose, rather than serving as an interactive data analytics tool.

Therefore, there is a need for the support to conceal the unnecessary technical complexity of creating a dynamic model from the data analysts, while allowing the data analysts to focus on the semantic level of the situation modeling. Such support is expected not only to reduce the time and efforts required for the modeling, but also to be able to empower more data analysts or even non-technical personnel to be capable of constructing a dynamic situation model.

> **Design consideration:** Support the users in constructing an interactive, dynamic, and computation-friendly situation model.

4.4.3 Major Component 3: Prognostic Insight

Prognostic insight is the prediction of the future state of the problem situation and the assessment of the impacts of possible actions. Prognostic insight provides users with the knowledge necessary to decide on the most favorable course of actions to meet their objectives. It happens during the knowledge actualization phase of data analytics. At this phase, the users already have a big picture about the

problem situation; they are interested to know how the big picture would change within the future timeline of interest, or how the big picture would change if certain actions were taken. Being actively and confidently aware of the plausible future, the users can proactively allocate resources to achieve their objectives. As a result, prognostic insight enables the users to adopt an anticipatory strategy to the problem situation with a faster cycle of insight creation, which is critical in complex and dynamic environments typical of modern organizations (Chen and Lee 2003).

Figure 31 shows how prognostic insight is achieved from synergic insights. Based on the users' current understanding of the problem situation, the users predict the future states of the problem situation. Multiple versions of the future might be generated, as to present different plausible developments or different interventions to the problem. Each of these versions is thus a hypothesized scenario. Subsequently, these different hypothesized scenarios are assessed in the light of the users' objectives and constraints.

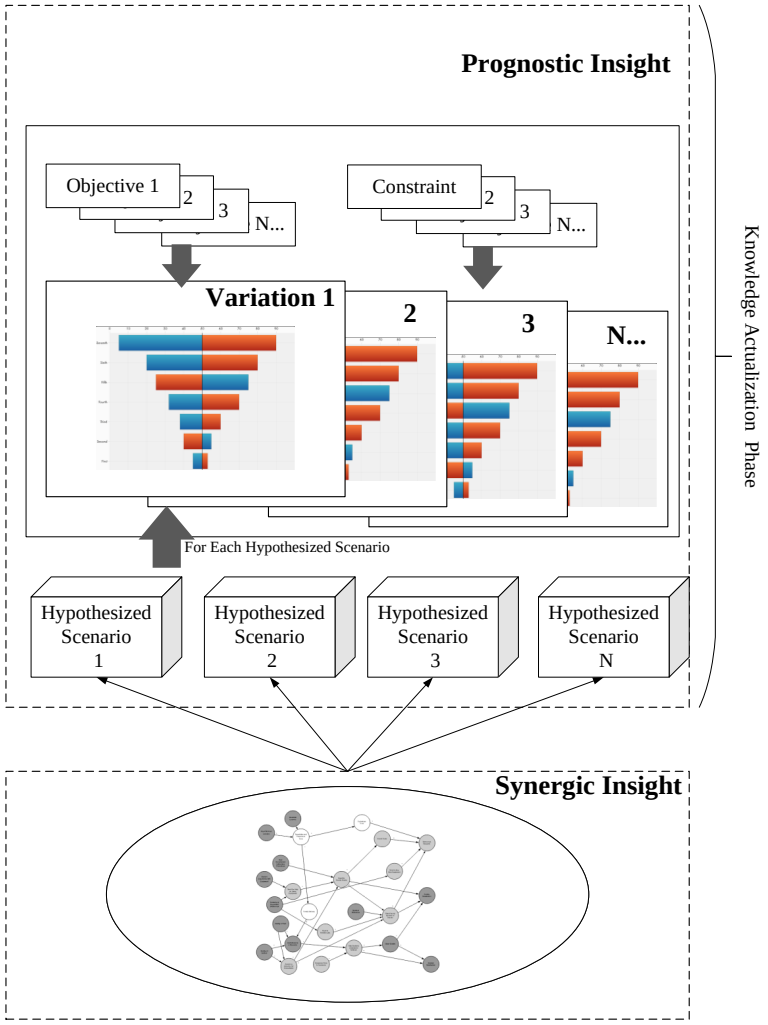


Figure 31. Prognostic insight: predictive + prescriptive insights

An example of prognostic insight in a stock market analysis would be the analysts gained understanding about how the market landscape will be in next year. Based on this big picture of the

future, the analysts then predict the influences of the stock market conditions on a number of stocks that they think would perform favorably. At this point, they might be confidently expecting that four of the stock prices will rise between 7 to 15%. Thus, these four stocks are the potential courses of action. However, at this point, the analysts do not know how the limited capital should be allocated to meet the objectives. The objectives are conflicting in nature: precisely, the analysts need to achieve a minimum of 9% annual growth, while keeping the risk below 25% of the fund. Added to the complication is the uncertain nature of the complex problem. The big picture of future predicted by the analysts is subject to the various uncertainties which may cause the actual price to swing. Therefore, the challenge for the analysts is to find out the optimal fund allocation that can best meet the objectives yet be less susceptible to the uncertainties.

4.4.3.1 Predictive Insight

Numerous research projects have pointed out that users engage in hypothesis generation and validation process in their mind during an analytical task (Keim, 2002; Lipford, Stukes, Wenwen, Hawkins, & Chang, 2010; Pohl et al., 2012; Yedendra B. Shrinivasan, 2010). After users build the situation model, users often interact with the situation model to gain deeper and dynamic understanding about the situation model. For instance, they want to perform a variety of "what-if" analyses based on the model. This allows them to test their speculations of what is going to happen and how it is going to influence the overall problem landscape. For instance, based on the understanding of how macro-economic factors influence different industries, which in turn influences the stock price, the users want to know what will happen if the unemployment rate increases from 2.5% to 5%, or what if the electronic export tax increases by 12%, or the combination of both. The outcome of these deep interactions between the model and the users changes the users' understanding of how the problem situation works and compels them to restructure their mental schemata (Weick, 1995). Such activity allows the users to gain deeper insight into the future states of problem situation.

Predictive insight is achieved when the users have understood the future states of the problem situation. *Figure 32* shows how predictive insight is derived from a situation model. The process requires users to engage in mental simulation and scenarios building. Based on the understanding of the complete problem landscape, users develop one or more hypothesized scenarios. There are two types of hypothesized scenario can be developed. The first type attempts to identify the most likely explanation to the situation without intervention by using different possible explanations to act as the competing hypotheses. The second type of hypothesis involves intervention, in which different possible courses of action are introduced into the situation model. In this case, each course of action results in one hypothesized scenario. Different courses of action are acting as the competing hypotheses.

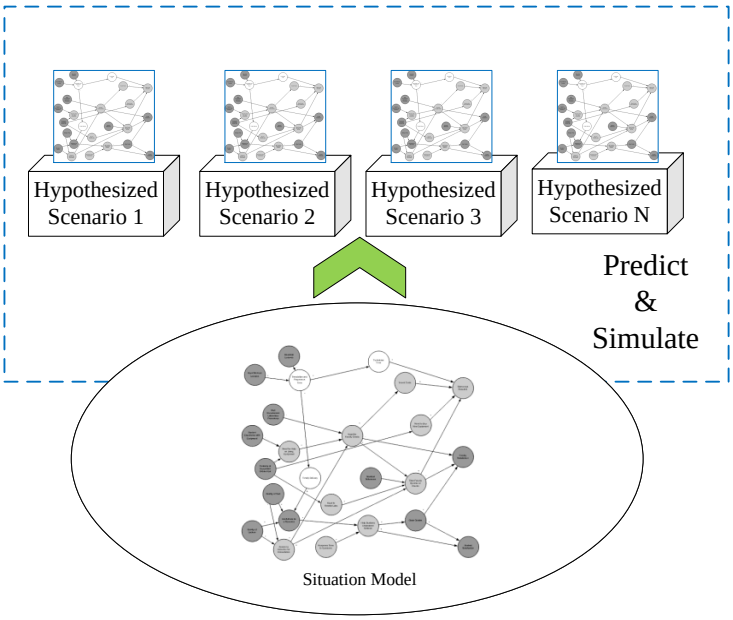


Figure 32. Predictive insight involves predict & simulate

Users commonly develop more than one hypothesized scenario, either consciously or subconsciously. Different hypothesized scenarios might be different due to the variations in the predicted situation model. The variations can be at the parameter level, the argument level, or the model level, as described in *Table 8*. Although it is easy to distinguish these variations conceptually, that is not the case in practice. A single hypothesized scenario might contain several variations of the same level and at the same time variations from different levels. As a result, the complexity of a hypothesized scenario can grow exponentially and the number of hypothesized scenarios may go beyond being manageable as the variation of scenarios increase.

Table 8. Variations in hypothesized scenarios

Variation level	Description
Parameter level	• Variations that involve one or more parameters of an argument • The structure of the situation model remains unchanged. • Example: - To compare the future market conditions of 7% and 9% of interest rate on short-term loan.
Argument level	• Variations that involve one or more arguments • Involves a change in the structure of the situation model, such as adding and removing an argument.

	• Example: - To compare what would happen if the country's GDP had influence directly on the stock price, rather than intermediated through the consumer price index.
Model level	• Variations that involves most arguments and causalities in the model • Huge change in the structure to the point where two hypothesized scenarios have very little similarity in terms of the arguments. • This often implies the change in the perspective of how the users understand the problem. • Example: - Two different hypothesized scenarios from the same users. The structure was informed by two different theories of how the market works. - Two different hypothesized scenarios from two users who have very different perspectives of how the stock market work

The prediction and simulation go hand-in-hand to analyze a hypothesized scenario. Recall that each hypothesized scenario is based on the users' situation model: that is, the overall understanding of how different actors, entities, events, and factors have influence on each other. Structurally, it is a network of arguments (represented by the nodes) and causalities (represented by the arcs). The users often start off the prediction of how the states of these arguments would be in the future. For instance, the users would predict how the inflation rate will be. The prediction could also be the result of different speculation of "what if" scenarios. Then, the users try to mentally simulate to see how the inflation rate will influence other components of the problem situation, based on their understanding of the connection between the components.

As cognitive activities, the mental prediction and simulation conventionally occur entirely in the analysts' mind. However, reasoning about hypothesized scenarios imposes exponential costs on the user's cognition (Pirolli & Card, 2005). When the cognitive resources are used to "hold" a hypothesized scenario, there will often be insufficient cognitive resources to predict or simulate the scenario. Humans are generally poor at simulating a "model". Mental simulation becomes nearly impossible when there are many interrelationships, which are common in complex analytics problem. Not mention about developing and analyzing multiple hypothesized scenarios. Researchers have shown that human cannot reason effectively about scenarios that are unavailable to them (Chinchor & Pike, 2009; Heuer, 1999). It is also well proven that doing complex analysis primarily in one's head is more prone to various cognitive biases (Thomas et al., 1993). On the other hand, explicit analytical reasoning helps to ensure more rigorous thinking, thus reducing the impacts of some cognitive biases. Therefore, one leverage

point is to improve the users' capacity to attend to more of the structure of the hypothesized scenarios, by externalizing the hypothesized scenario from the user's mind to the system.

> **Design requirement:** Support the modelling, representation, and storage of hypothesized scenarios

The problem is not solved, even with the support to build, holding, and visualize the hypothesized scenarios. The real challenge is to enhance the accuracy of the mental prediction and simulation. One study has shown that 90% of people failed to accurately simulate a model in their mind, even a simple one (Richmond & Peterson, 2001). There are long and well established predictive analysis techniques, such as linear regression and logistic regression. These traditional prediction techniques have some limitations. Firstly, they are static and rigid processes. Secondly, these results are often deterministic predictions that do not fit the highly uncertain nature of the complex analytics problem.

Traditional prediction techniques are rigid. The predictive algorithms are rigid in the sense that they use standardized formula to produce the result, so it is often not possible for the users to flexibly tweak the algorithms to reflect their own understanding or knowledge. The techniques also require all data for the prediction to already be made available in the system. In practice, however, missing data is common and often the users' subjective inputs are important for filling the data gaps. Without a way to allow the users to steer the prediction and simulation, the accumulated knowledge and experience of the expert users over years are neglected. An expert user with 25 years of experience and a novice user without experience may use the same prediction algorithms in a similar way, as long as they know how to use the analysis tools and understand the concepts that underpin the techniques.

Moreover, these predictive techniques are also deterministic in nature (Zuk & Carpendale, 2007). In other words, these techniques do not deal with uncertainty and incompleteness in the data. However, complex analytics tasks often require stochastics prediction rather than a deterministic one. This requires the prediction and simulation to be done in terms of probabilistic analysis. Humans have been proven to be weak at probabilistic analysis without external mathematical aids (Keim et al., 2010). For instance, availability bias and representativeness bias often lead to an excessively inflated prediction of the occurrence of unlikely event.

Therefore, this study asserts that it is critical to aid the prediction and simulation with quantitative techniques to enhance the accuracy of hypothesized scenarios. There are some predictive analytics techniques, such as machine learning and computer-aided reasoning, that can be incorporated to augment the reasoning process. Such structured reasoning encourages rigorous and logical processing, which will enhance the validity of the reasoning outcomes and reduce cognitive pitfalls. Research has also shown that users can benefit in such complex problem situations by engaging in gaming-like

processes to clarify the nature of the problem (Mirel 2004). In other words, this requires a blend of experimental and analytics techniques to deal with the uncertainty and the dynamic nature of the problem model.

> **Design requirement:** Support prediction and simulation with the aid of computer-aided reasoning that can be flexibly steered by the users to reflect their intention, judgment, and knowledge.

4.4.3.2 Prescriptive Insight

To this point, the users might have identified the most plausible scenarios or potential courses of action. But they are not yet confident which of the courses of action can best meet their objective, within the constraints they have, and how these courses of action would react to the uncertainty in the scenario. In practice, given that resources are often limited and scarce, the users have to choose among the different options in order to act effectively and efficiently (Klein, 1993).

Acting effectively and efficiently implies the importance of the optimal allocation of resources. Note that optimization of resource allocation was not part of sensemaking theory, nor was it implied by situation awareness theory. However, resource allocation is often an inevitable process in solving complex problem. For example, stock investment analysis involves allocating the limited fund to different stock options; environmental policy involves allocating limited money and human resources to different projects; and disaster damage control is similar. The resource allocation depends highly on how the hypothesized scenario may vary from the expectation due to uncertainties. Therefore, optimization is the key for achieving a resource allocation plan that can maximize the objectives, given the constraints, while compensating for the uncertainties. Therefore, this study includes optimization as the additional concept, beyond what the sensemaking and situation awareness theory have included.

Prescriptive insight is achieved when the users are aware of what the optimal resource allocation is that will enable them to maximize the objectives, while compensating for the risks caused by the uncertainty inherited in the hypothesized scenario. As the result, prescriptive insight enables the users to understand the potential outcomes of both their action and the associated risks. More importantly, the insight allows them have an optimal plan that is ready to be translated to a solution for the analytics problem: that is, allowing them to act on the insight.

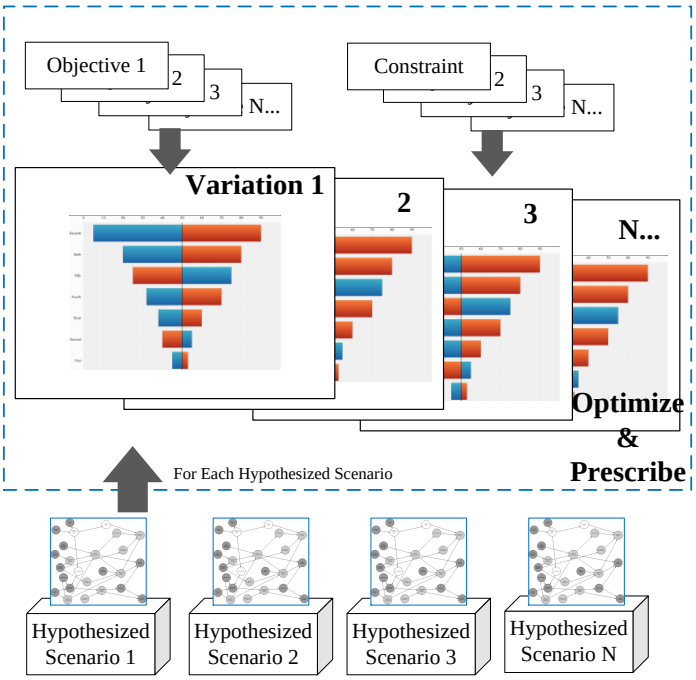


Figure 33. Prescriptive insight involving optimization and prescription

Figure 33 illustrates the activities involved in deriving prescriptive insight. Previously, multiple hypothesized scenarios might have been developed by the users. After analyzed the scenarios, they would often narrow these down to a handful of the most plausible scenarios. At this stage, each of these hypothesized scenario is scrutinized by being evaluated against the users' objectives and constraints, while factoring in the uncertainty inherit in the hypothesized scenario. Explicit consideration of objectives, constraints, and uncertainty is vital for complex problem solving (Thomas & Cook, 2005). In practice, analysts often experiment with different combinations of objectives, constraints, and uncertainty. As a result, multiple variations of the hypothesized scenario are formulated. *Table 9* provides detailed descriptions of the elements contributing to the variation.

Table 9. Possible elements of a variation

Elements of Variation	Description
Objective	• Objectives in the complex problem are not static and rigid. Often, users may revise their objectives as they progress through the data analytics. • Objectives in a complex problem can be more accurately described as maximization or minimization, rather than the achievement of a specific deterministic objective value. • For example: Users try to assess the impacts of the hypothesized scenario on other objectives and constraints,

	○ when changing the minimum ROI from 9% to 14.5% (an example of maximization objective). ○ when changing the maximum investment risk from 25% to 50% (an example of minimization objective).
Constraint	• Similar to objectives, constraints are dynamic and flexible. May be revised as the users discover the constraint is not realistic, or if they believe they can convince the client to provide more resources, given the scenario that they discovered. • Constraints can be applied to resources, rules, or preferences. They can often be defined as a threshold value. • For example: ○ Resources: reducing the investment capital from $500,000 to $300,000 given the impression of bearish market. ○ Rule: Capital goes to "Consumer Electronic" stocks must not more than 25%. ○ Preference: A minimal of 60% of capital shall be invested to Apple Inc.
Uncertainty in the Scenario	• The uncertainty is inherited in the structure of the hypothesized scenario. • It depends on how certain / confident the users are in the arguments or causalities within the hypothesized scenarios. • For example: ○ How confident that the inflation rate in next year will fall between 3 to 4%. ○ How confident that the increment in the inflation rate will be associated the increment in consumer price index.

In practice, the variations of the hypothesized scenario that the users consider can be very complex and "messy". A variation may involve multiple changes in the objectives, resources, rules, and preferences, and the uncertainty at the same time. As a result, a massive number of variations of a single hypothesized scenario need to be considered by the users. Therefore, by having several hypothesized scenarios can increase the variations exponentially.

Given the complexity a variation of the hypothesized scenario can have, it is very difficult for human users to accurately assess the impacts of different variations. Additionally, the human analysts are not good at objectively gauging the risk caused by the uncertainties, without external mathematical aids. Human analysts have the tendency to overinflate their confidence in the conclusion that they derive,

based on vague and subjective assessments. Reliance on human subjective judgment of the risk is highly risky, especially when dealing with an unfamiliar problem situation where the stake is high. To make informed decisions, it is critical for data analysts to know what is the chance of things do not go accordingly to plan, and how that is going to reflect on the objectives and constraints.

Therefore, this study asserts that this is the leverage point where the computer-aided risk assessment techniques can be adopted to enhance the rigor and accuracy of the risk assessment in the complex analytics problem. This assertion is aligned with the fact that the use of machine processing and visualization, to complement the human analyst's abilities in understanding uncertainties is listed as one of the important research agenda items in the visual analytics field (Keim et al., 2010). Computer-aided risk assessment enables the analysts to experiment with a large number of variations within a hypothesized scenario efficiently, while offloading the analysts to focus on value-added high-level reasoning. More importantly, it allows the analysts to quickly understand the impacts of the variations and to assess the associated risks. Without the computer-aided risk assessment, even the best-built hypothesized scenario can be tested and improved only by relying on the feedback through the real world. Richmond and Peterson (2001) described the real-world feedback as "*very slow and often rendered ineffective by dynamic complexity, time delays, inadequate and ambiguous feedback, poor reasoning skills, defensive reactions, and the costs of experimentation*".

> **Design Requirement:** Support users in accurately and rigorously assessing the risks associated with the courses of action.

Moreover, the challenges for the human analysts are not just to understand the potential risks, but also to optimize the resource allocations, in order to simultaneously fulfill their conflicting objectives, within the boundary of their constraints, while compensating for the risks. Given the complexity, it is nearly impossible for human analysts to do the mental calculation for deriving the optimal resource allocation. Therefore, they often rely on intuition for resource allocation. Worse, they have the tendency to allocate the resource evenly across the few options that they have. Therefore, without support, it becomes all too easy for the resource allocation to be driven by unconscious bias. Therefore, this is an obvious leverage point where support can be provided to enhance the user's analytical performance.

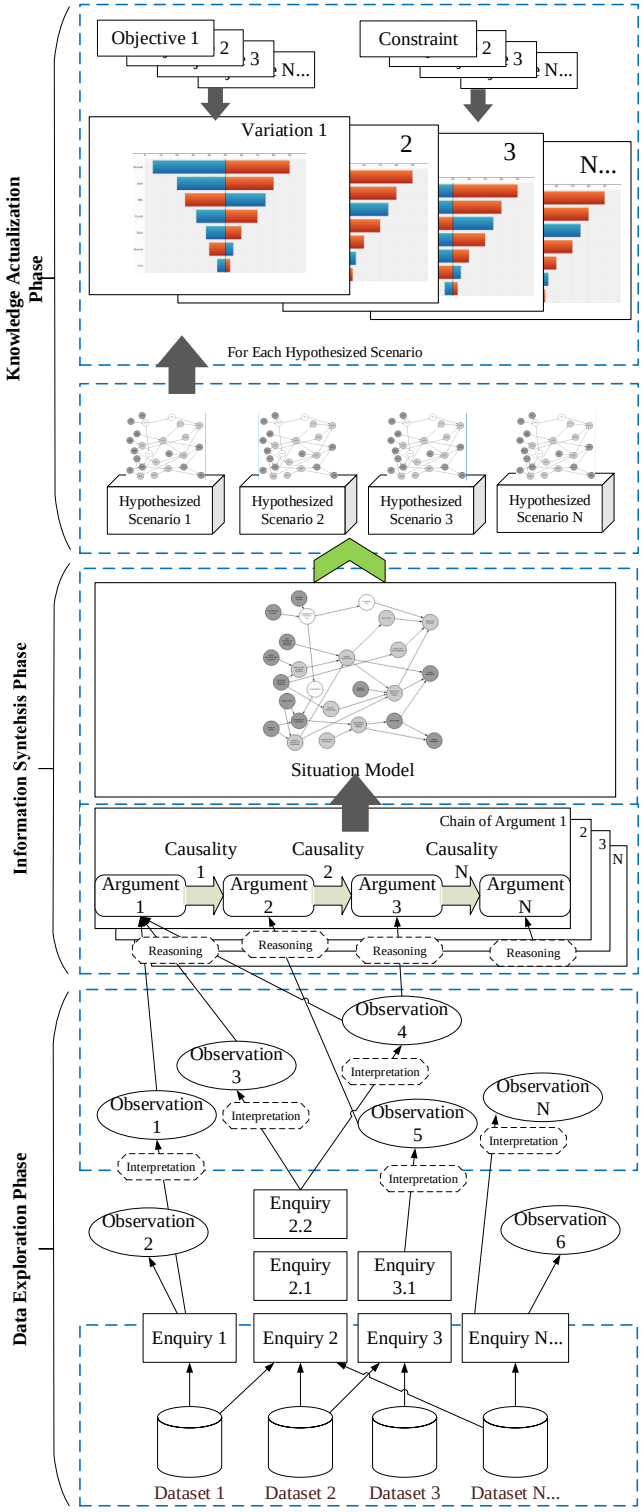
Nevertheless, the real challenge to deliver such support in complex analytical tasks is that the optimization needs to take consideration of the uncertainty. For example, given that the selling prices of product A and B are \$15 and \$19 and the costs of producing product A and B are \$6 and \$7, the algorithm can optimize how many product A and B should be produced if \$5000 of capital is available. As illustrated in the example, the costs and prices are certain. But in complex analytics problems, all these numbers can be uncertain. For example, the selling prices and costs of product A and B can fluctuate by 50% and 70% respective. Then, the purpose of optimization is to seek for the balance points

that would minimize the negative consequences, given all the uncertainties that can potentially happen. This requires stochastic optimization, rather than the conventional deterministic optimization. However, achieving stochastic optimization often requires specific knowledge and skill. For instance, it may require specific programming-like syntax to specify the problem. A significant proportion of users do not possess the required knowledge and skill, which prevents them for taking advantage of such techniques.

> **Design Requirements:** To support users in optimizing the resource allocation that can meet the conflicting objectives within their constraints, while compensating for the risks.

---- This Space is Intentionally Left Blank -----

4.4.4 Summary of the Hierarchical Framework of Insight



Outcome: Prescriptive Insight
Activities: Optimize & Prescribe

- To understand the impacts of a hypothesized scenario on the objectives and constraints
- To aware of the optimal resource allocation plan to maximize the objectives within the constraints, while compensating for the potential risks caused by the uncertainty inherited in the scenario
- Enable the users to confidently understand of the risks of their planned courses of action, and to have optimal plan that is readily to be translate to action.

Outcome: Predictive Insight
Activities: Predict & Simulate

- To theorize the states of the situation model in the future or after intervention is introduced.
- Enable the users to gain deeper insight into the plausible scenarios

Outcome: Comprehensive Insight
Activities: Connect & Build

- To develop a closed-system model that replicates the phenomena intended to be solved or comprehended
- Enable the users to understand the problem situation as a big picture and the how underlying components influence each other.

Outcome: Integrative Insight
Activities: Integrate & Synthesis

- To create high-level knowledge by integrating multiple low-level observations and synthesizing with users' implicit knowledge
- Enable the users to go beyond dry data to understanding the actors, entities, events, or concepts that are significant for solving the analytical problem

Outcome: Perceptive Insight
Activities: Perceive & Interpret

- To extract meaningful information from data by perceiving and interpreting the technical analysis results
- Enable the users to understand the low-level information about a very specific aspect of the overall problem.

Outcome: Identification Insight
Activities: Search & Filter

- To explore relevant data elements from massive amount of data
- Enable users to identify relevant data elements for the analytical problem.

Figure 34. Summary of HIVE framework

The three major insight components characterize different extents of abstractions, content granularity, objectivity, human reasoning, and domain value. *Figure 35* summarizes the characteristics of these three components.

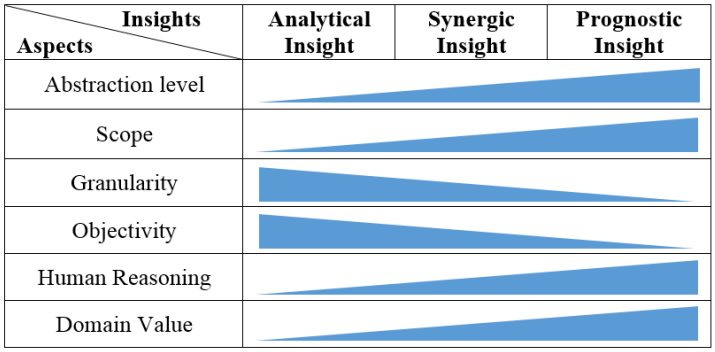


Figure 35. Characteristics of the three major insights

The level of abstraction and content granularity are reversely associated: the higher the abstraction level, the lower the content granularity. Prognostic insight at the top layer characterizes the highest level of abstraction and the widest scope, while it has the lowest level of content granularity, as indicated by the first three rows in *Figure 35*. A prognostic insight is usually concerned about interactions of various high-level factors in the entire problem situation, such as consumer purchasing trends and changes in an international trading policy. At this level, the granularity is relatively low, as the detailed data such as sales volume of different product lines grouped by regions is hidden for the reasoning and analysis at this level.

The objectivity of the insights is high at the lower level and low in the higher level. For instance, analytic insights at the lowest layer are mainly concerning quantitative and objective information such as technical indices produced by analysis tests. Subjectivity of the insight increases toward the higher end as more qualitative data is being integrated. Subjective information such as domain knowledge, contextual information, and judgmental heuristics become more important and prevalent toward the higher-level insights, such as synergic insight and prognostic insight.

The higher the insight located in the framework, the more important the subjective human reasoning is for deriving that insight. Information at the higher level often requires complex processing, such as synthesizing information, extracting semantic meanings, and generating hypotheses. Such qualitative reasoning is the weakness of conventional computational methods. Relatively, human reasoning and judgmental heuristic allow the users to perform effectively in deriving higher-level insights such as synergic and prognostic insights. In other words, the workloads on human reasoning increase toward the higher end of the insight layers.

Domain value tends to increase toward the higher layers. In other words, high-level insights such as prognostic insight are relative ready to be translated into practical decisions or actions. Insights on

the higher layers tend to incorporate more domain or context-specific information into the analysis, such as user's current objectives, constraints, and solution alternatives.

4.5 Summary of Design Requirements

Table 10 lists all the design requirements derived previously in accordance to their corresponding problem-solving activities and insights components. In the next chapter, these design requirements are used to formulate the design principles of this study. Together, the design principles form the overall design framework of this study.

Table 10. Summary of design requirements

Activities	Insights	Design Requirements
Search and Filter	Identification insight	• To support the users to effectively explore large number of data elements
Perceive and Interpret	Perceptive insight	• To support the users to capture and manage their observations, including the underlying interpretation. • To support users to create joint summary from their observations.
Integrate and Synthesize	Integrative insight	• To support the users to create, manage, and retrieve high-level knowledge based on low-level analytic insights and reasoning.
Connect and Build	Comprehensive insight	• To support the users to identify a preliminary structure of the situation model • To support both quantitative and qualitative information to build the situation model • To support the users in constructing interactive, dynamic, and computation-friendly situation models
Predict and Simulative	Predictive insight	• To support the modelling, representation, and storage of hypothesized scenarios • To support the prediction and simulation with the aids of computer-aided reasoning that can be flexible steered by the users to reflect their intention, judgment, and knowledge.
Optimize and Assess Risk	Prescriptive insight	• To support users in optimizing the resource allocation that can meet the conflicting objectives within their constraints, while compensating for the risks.

		• To support users in accurately and rigorously assessing the risks associated with the courses of action.

The commonality between design requirements is that they all require a balanced blend between 1) the flexibility of human knowledge and reasoning and 2) the rigor and efficiency of the machine-driven computation. Most, if not all, of these requirements cannot be fulfilled by either the human-oriented or the machine-oriented approach alone.

---- This Space is Intentionally Left Blank -----

Chapter 5
Developing the Conceptual Design Framework

5.1 Overview

The purpose of this chapter is to develop a conceptual design framework that can be used to inform the design and the implementation of data analytics systems which can effectively support the users' problem-solving activities in different phases of data analytics process. The conceptual design framework comprises a set of design principles, of which each design principle is formulated based on the design requirements from previous chapter. The design principles, in turn, are used to inform the design of the features and functionalities in a prototype system. This study theorizes that, together, the design principles provide a coherent set of supports to enhance the users' analytical performance in the problem-solving activities.

In relation to the research objective, this chapter addresses objective C -- To create a design for a data analytics system that supports the processes and requirement. Section 5.2 introduces the design philosophy that governs the overall design solution. Then, Section 5.3 presents the design principles, and also 1) explains the rationales behind the design, 2) describes how each of the design principles is used to inform the implementation of the system features, and 3) explains the hypothesized effects to be achieved by the design principles. Section 5.4 presents the design framework, a big picture of how the design principles work together.

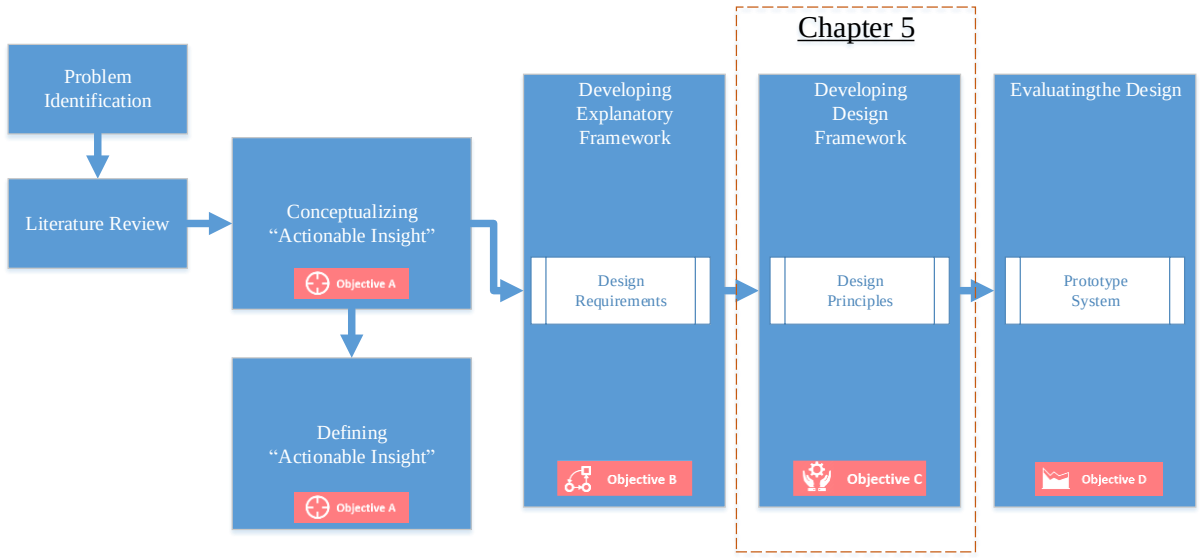


Figure 36 Contents of Chapter and Research Objective

5.2 Design Philosophy

The review of the design requirements revealed that they all require a balanced blend between 1) the flexibility of human knowledge and reasoning and 2) the rigor and efficiency of the machine-driven computation. In other words, the requirements cannot be fulfilled by either the human-centric or the machine-centric approach alone. In accordance with this commonality in the requirements, this study contends that the design solution should adopt a unified design theme in order to be coherent.

Design philosophy is the central design doctrine that guides a design work. The design philosophy of this study is called "*machine-augmented cognition*". As suggested by its name, the ideology of *machine-augmented cognition* is to amplify the human analytical reasoning capability with computer-aided techniques, with the goal of solving complex analytics problems. The computational aids can come in the forms of interactive visualizations, structural reasoning techniques, mathematical modelling, artificial intelligence, and other computational algorithms. A distinguishing aspect of this approach is that human cognition is the primary driving force in the analytics task. Human cognition steers the way that computational aids act, as the scaffolding or the catalysts, to boost the analytical performance of the human data analysts. This study believes that data analytics systems which are developed based on the analysts' cognitive orientation will be able to enhance their performance in solving complex analytics problems.

The main motivation this study chooses for this design philosophy is the current limitations of pure machine-centric and pure human-centric approaches in solving complex analytic problems. Although machine learning has been the data analytics approach under the spotlight for 2 to 3 years due to the active involvement of big players such as Google and IBM, this study asserts that machine learning is not a silver bullet to all types of analytics problem, particularly complex analytics problems. As data mining innovator Sankar (2013) explained, a complex analytic problem such as assessing the impacts of an international policy or identifying a terrorist network is not about finding or creating powerful algorithms, but is rather about finding or creating the right symbiotic interactions between the data, the computations, and human cognition. Human cognition should take the lead in complex analytic problems since only the human analysts can determine the contexts, meanings, and relations of the discoveries made (Meyer et al., 2010). Additionally, heuristic judgment is required to make the best possible evaluation of incomplete, inconsistent, and potentially deceptive information. However, approaches that rely solely on the human analysts to solve complex analytics problems have become known for their inefficiency and their erroneous results. Although it is scarce, there is an increasing amount of research advocating for a human-machine symbiosis approach for solving complex problems. These human-machine symbiosis approaches have been recognized as a means to solve complex problems in many important fields (Thomas & Cook, 2005).

Figure 37 shows the relative a position of machine-augmented cognition approach in relation to other data analytic approaches on a continuum. The left end of the continuum denotes fully manual analytic; the right end denotes fully automated analytic. The closer an analytic approach is located towards the left end of the continuum, the more intensive the human-based reasoning required. In contrast, the further towards the right end of the continuum an analytic approach is, the more readily it can be fully automated without human intervention.

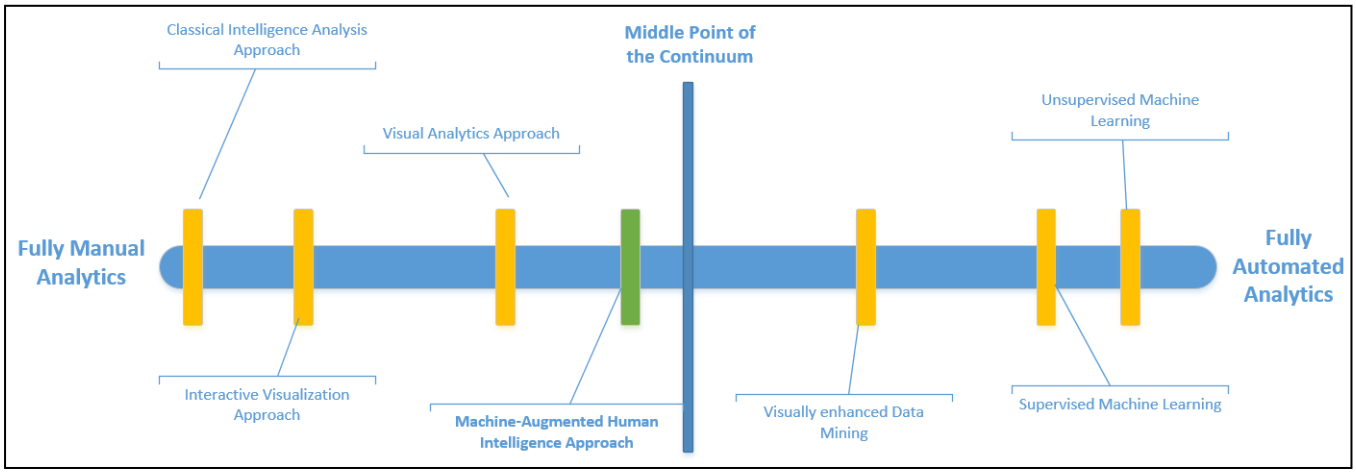


Figure 37. Relative position of "machine-augmented cognition" approach

Machine-augmented cognition is located slightly to the left side of the continuum's middle point. In this approach, human analysts prime the data analytic by proactively providing the contextual information and subjective inputs, deciding and fine-tuning the algorithms, and overwriting the computations with their logic and reasoning. The computational aids are to enhance their analytical reasoning process, rather than to automate the analytical reasoning. The enhancement includes ensuring the rigor of the reasoning, minimizing the cognitive bias, and reducing the cognitive loads. Humans and machines work in a complementary manner to reinforce each other's strengths while counteracting the weaknesses. The goal of machine-augmented cognition is to enhance the human-information discourse in order to achieve better human analytical performance along the problem-solving activities.

Compared to its neighbor approaches, *visual-enhanced data mining* is the approach in which data mining is the primary data analysis means (Bertini & Lalanne, 2009). Human reasoning is mainly required for interpreting the visualized results. To the left, in the *visual analytics* approach, human is the primary means of visually discovering insights from interactive visualizations. Computations are mainly used for data reduction to facilitate the visual discovery, rather than supporting the human reasoning process.

In this study, the goal of the "machine-augmented cognition" design philosophy is to provide a unified theme to the design principles. Following the same design philosophy, all the design principles should address the design requirements with solutions that seamlessly integrate both the computational

aids and the user-driven analytical reasoning. This study contends that the *machine-augmented cognition* design philosophy can produce designs that enable the data analysts to effectively perform the problem-solving activities required to achieve actionable insight.

5.3 Design Principles and Operationalization

The conceptual design framework consists of a set of design principles. These design principles are manifested as one or more features in a prototype system. The objective of this section is to present the design principles and their operationalization from the conceptual design to the tangible information system functionalities in the prototype system. Each of the design principles is described in the following schema of design overview, IS initiative, mechanism, and intended effects. *Table 11* further elaborates the content of these components.

Table 11. Components of a design principle

Components	Descriptions
• Design Overview	• Recapitulates the design requirements • Provides a brief overview of the design principle
• IS Initiative (I)	• Provides the conceptual discussion on how the information system supports can address the design requirements. • Presents and explains the conceptual design.
• Mechanism (M)	• Describes the system functionalities that are actualized from the conceptual design. • Describes how the design principles are implemented in the prototype system.
• Intended Effects (E)	• Describes the conjectured effects of the design initiatives. The discussion includes how the effects could manifest in the analyst's behaviors or performance.

The conceptual design proposed in this study is developed for complex analytics problems in general, and can be applied to most complex analytics problems. In order to evaluate the design, a prototype system was developed to operationalize the conceptual design into tangible system features. This study contextualizes the prototype development and testing in the stock market investment domain. This implies that some of the system features derived from the design principles are specific to stock

market investment. Four reasons emerge for the decision to choose stock market investment as the domain.

- Firstly, stock market investment is a suitable scenario to emulate a complex analytics problem. It involves a large number of interconnected data elements, it requires human knowledge and judgment to solve the problem, and it contains multiple conflicting objectives.

- Secondly, for evaluating the design in this study, an analytical task with reasonable duration can be designed. A stock market investment task can be reasonably scaled down to one to two hours without significantly altering its complexity or the nature of the task.

- Thirdly, stock market investment is a relatively common domain which can be learnt or understood by most laypersons, without extensive, specialized training. This increases the participant pool that this study can recruit for the evaluation.

- Fourthly, the data used in the analytics systems can be acquired easily and in large scale. The data also do not involve sensitive information as opposed to other domains such as crime investigation. More importantly, historical stock market data from the real world can be used. This enables the designed task to be realistic, rather than using synthetic data.

5.3.1 Enabling Divergent Exploration

5.3.1.1 Overview of Design Principle

Design Requirement: To support effective exploration of a large number of data elements.

Divergent exploration is a design principle that advocates for the importance of diverse exploration during the data exploration phase. The objective of the design principle is to encourage users to explore the data from diverse perspectives. The design principle consists of two IS initiatives, namely 1) enabling data divergence and 2) enabling enquiry divergence. The former enforces heterogeneity in the data sources, whereas the latter enforces heterogeneity in enquiries. This study conjectures that divergent exploration can reduce the chances of premature exclusion of certain data elements based on the analysts' preconception of the problem. The design principle could aid the users in the search & filter activity.

5.3.1.2 IS Initiatives

The design principle of "divergent exploration" advocates for the enforcement of multi-perspective views of the data elements. Explanations of the two specific design initiatives follow.

Divergence at the data layer is especially critical for complex analytics problem for several reasons. Firstly, due to the unique and unpredictable characteristics of the problem situation, the data that initially seem to be logically irrelevant potentially contain vital clues to the problem situation (Lefebvre, 2004). Some data elements which seem to be irrelevant are often intertwined with the key information that is of interest to the analysts. These data can be useful in two ways: 1) to predict the key information if the key information is incomplete, 2) to assess the reliability of the available key information. In complex problem situations where data are often imperfect, divergence at the data layer provides a means of enhancing the quality of the data sources. Moreover, data heterogeneity improves the accuracy and completeness of the data elements that used to inform the analysts' problem assessment and solution, thus leading to higher quality solutions. Solutions to complex problems often need to satisfy multiple conflicting objectives where each represents the vested interest of different stakeholders. Divergent data would help the analysts to develop solutions that are well-rounded from multiple perspectives and that would reduce the blind spots of the solutions. Pirolli and Card (2005) have shown that expert analysts often set their filters for information lower, thereby accepting more irrelevant information because they want to make sure that they do not miss something that is relevant.

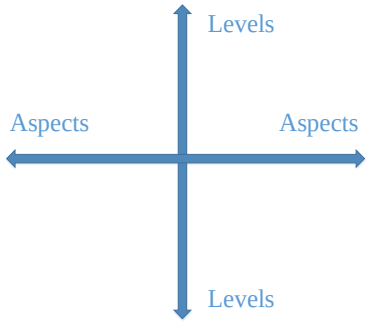


Figure 38. Dimensions of data divergence: aspects and levels

At the data layer, data analytics systems should enable data elements from all different aspects and levels to be included and considered, from the early stage of the analysis and at any point of the analysis. *Figure 38* illustrates the aspect and levels as different dimensions of data divergence. Aspects refer to the horizontal expansion, which is centered against a topic, but which provides different angles of understanding of the topic. For instance, different aspects of stock options include the stock price, the company's profitability, the board capability, and the media appearance. On the other hand, "levels" refers to the vertical expansion, which involves factors at higher or lower abstraction level. For example, the country's macroeconomic conditions, sector-specific indexes, or industrial-wide factors provide different levels for the data divergence. The divergent data often imply the need for running data exploration across multiple datasets which sit in their own silo. Therefore, this study proposes a design in which the system should support users to easily integrate their datasets of choice, in order to create a centralized data source for generating their enquiries. *Figure 39* shows the conceptual illustration of this design.

Design: Supporting multiple datasets to be integrated
to create a centralized data source for enquiry generation

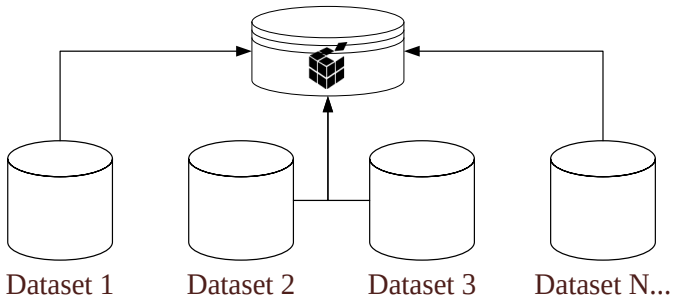


Figure 39. Supporting data divergence with dataset integration

Divergence at the enquiry layer is important for complex problem, as the analysts need to assess the problem situation from multiple perspectives (Jonassen, 2000). The ability to do so is relied highly on the analysts' ability to restructure or reframe the data elements on the fly to reflect their preliminary epistemic belief about the nature of the problem situation. Divergent enquiries would allow the restructuring and reframing of the data elements more intuitive, thus enhance the way data analysts perceive information from the data. This is because humans naturally understand information in the form of multidimensional. Therefore, if the data are represented in multiple perspectives, with each turned to a particularly important aspect of the data attributes, this could help the analysts to be more effective in discovering implicit information from the relationships between the data attributes (Green, Ribarsky, & Fisher, 2008; Hetzler, Whitney, Martucci, & Thomas, 1998).

At the enquiry layer, divergent exploration is achieved by multimodal enquiries. The purpose of multimodal enquiry is to allow the analyst to perceive the multi-facets of the data. In the context of visualization, multimodal enquiry involves simultaneous interactions with multiple coordinated visualizations. For instance, the time-series view (e.g. trend line) and the cross-attribute view (e.g. scatter plots) are coordinated in the way that when users choose a particular time period in the time-series view, the cross-attribute view will show only the data corresponding to that time period. Such enquiries allow the relations or patterns between multiple data perspective to be uncovered. Such an implicit information is implicit characteristic of the data that analysts derive from their rich interaction with the multimodal enquiry. The ability to discover implicit findings from data is critically important in the complex problem, where the key information is often not explicitly observable from the data. Therefore, this study proposes that the design should support users to generate multimodal enquiries. *Figure 40* shows the conceptual illustration of this design.

Design: Supporting the generation of multimodal enquiries
for perceiving multi facets of the data

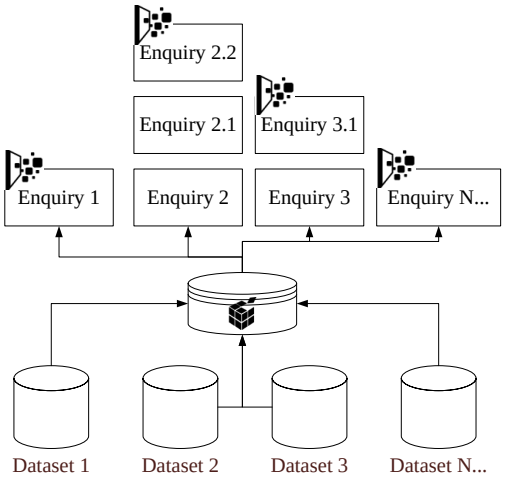
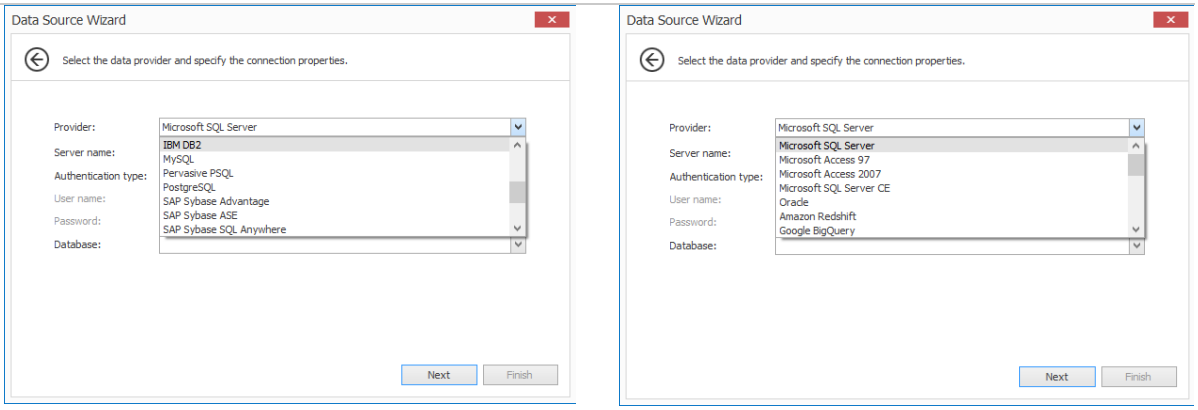


Figure 40. Support divergence enquiry with interactive multi-modal enquiries

5.3.1.3 Mechanism

The design principle is actualized based on these two design initiatives. Following the design "***supporting multiple datasets to be integrated to create a centralized data source for enquiry generation***", a data integration mechanism is implemented to enable the analysts to easily pull data from different repositories into a single data analytics project. It is common that data required for the data analytics sit on different repositories. For example, the operational data of an institution are often stored in a centralized database of the enterprise systems such as SAP, while ad-hoc data are often stored in the spreadsheet format. The mechanism allows users to connect to different data repositories with minimal technical knowledge about database connection. *Figure 41* shows the interface for select and connect to different common data repositories.



Data Providers from Enterprise-level Systems | Data Providers ranged from personal to cloud

Figure 41. Interfaces for pulling data from multiple repositories

After users have integrated the data from the different repositories, users can use the graphical interface shown in *Figure 42* to specify how they wish to link the data. *Box 1* shows the data tables from all the connected repositories, while *Box 2* shows the data elements of a selected data table. The users can specify new relationships between the data which originally were not specified in the repositories. This allows them to logically link the data as they need for their enquiries later. *Box 3* shows the automatically generated codes based on user interactions for the data selection, linking, and filtering, allowing advanced users to edit the code directly. Starting from this point, all the chosen data is loaded into the computer's random access memory (RAM). All computations and analytics will directly perform against the data stored in the RAM in order to enhance the computation process.

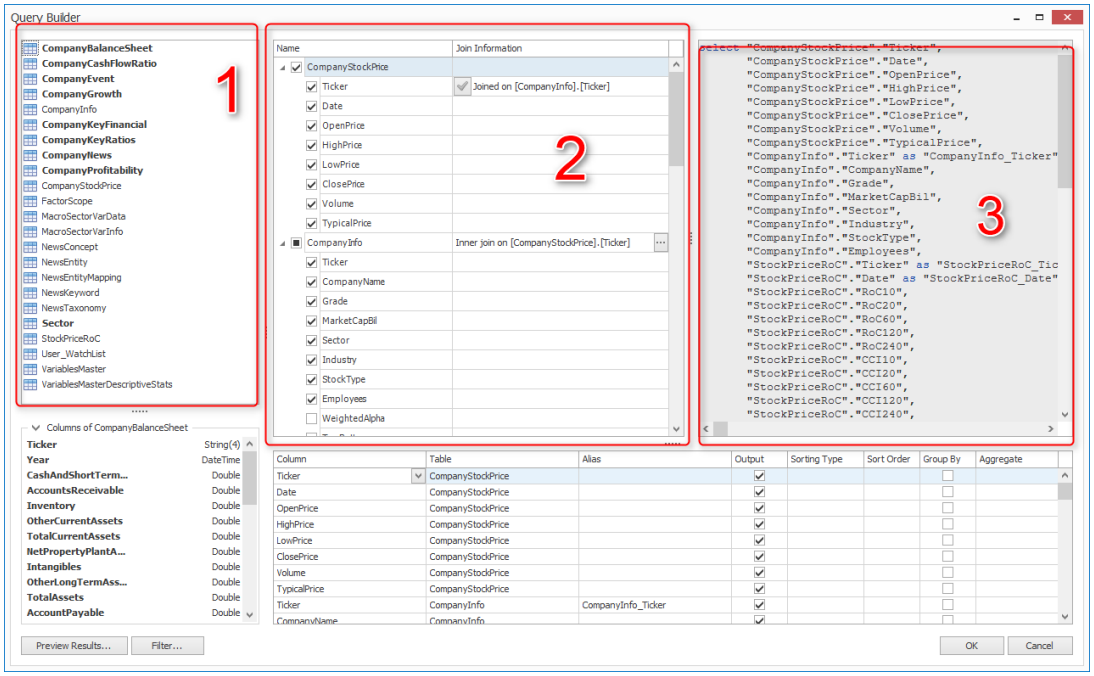


Figure 42. Enable users to integrate and link data with no technical skills required

Following the design "***Supporting the generation of multimodal enquiries for perceiving multi facets of the data***", a multimodal enquiry mechanism is developed. The users can use it to generate enquiries based on the data elements integrated. This study chooses to implement the multimodal enquiries in the form of visual analytics. Visual analytics is a type of data analytics approach that allows users to visually build and analyze data through interactive visualizations (*Extra: Why this Study uses a Visual Analytics approach for the Data Exploration Phase*", 105.

The visual analytics approach allows the users to easily build custom multimodal visual enquiries. The enquiries can be created through drag-and-drop, wizard, and interactions which require minimal technical knowledge of data visualization or manipulation. As a result, this enables analysts to intuitively create virtually unlimited combinations of multimodal enquiries. The design is in line with the assertion from Meyer et al. (2010) that, for data analysts to gain understanding of complex

information collections, they must be able to visualize and explore multiple facets of the information with ease. *Figure 43* shows the screenshots of the user interface which users can use to build custom multimodal enquiries.

Box 1 in *Figure 43*, which shows the previously integrated data elements, is now available to be used for generating the enquiries. To create visual enquiries, users just need to drag and drop the data elements into the corresponding fields in Box 3. There are different types of visualization that the users can choose from Box 2. Box 4 displays the visual enquiry, which can consist of one or more visualizations. The visualizations facilitate deeper understanding of the data through user interactions such as selection, filter, zoom in and out, annotation, and drill-down. Moreover, the visualizations within the same visual enquiry are coordinated. For example, in Box 4, the stock price (i.e. the line chart on top right) will be filtered according to stock that the users select on the table (i.e. the data grid on the left side) or the time period that users select on the time slider (i.e. the range control below the stock price). Such dynamic interaction with multiple visualizations is the main enabler of the multimodal enquiry.

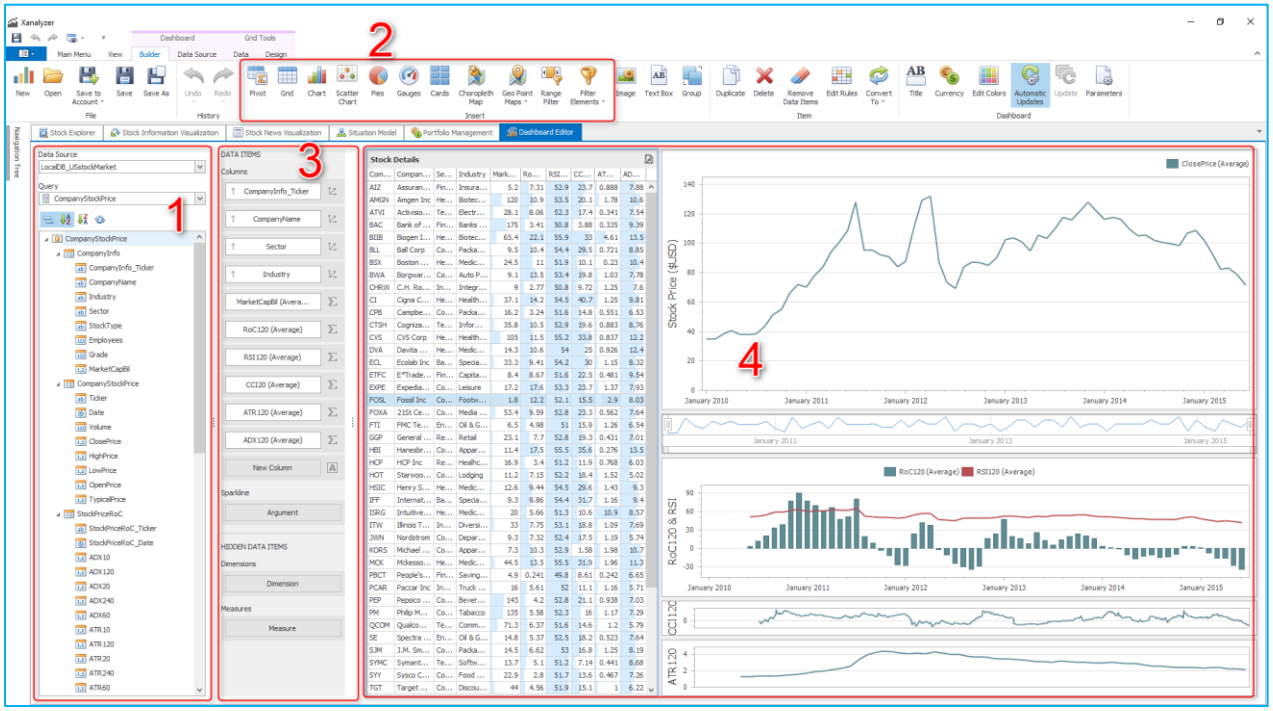


Figure 43. Multi-modal enquiry enabled by visual analytics

In addition, the users can flexibly rearrange the layout by grouping the visualizations under the tabbed window or distributing them across separate windows. This allows users to flexibly create multimodal enquiries on multiple subjects and have them displayed simultaneously across multiple monitor displays. Such design indirectly supports diverge exploration: making the enquiries visible to the users in single view could enhance their capacity to perceive and interpret the enquiries. The user

experiment conducted by Weick (1995) has shown that approximately 35 percent of the errors made in perceiving and understanding the information that is intended to be presented by the system are not visible to the users (e.g. there are other windows that overlay on top on each other and require constant switching between different windows). Therefore, the design here is believed to optimize the experience of analysts in perceiving and understanding diverse enquiries. *Figure 44* shows the ability to rearrange the tabbed windows to be viewed in parallel.

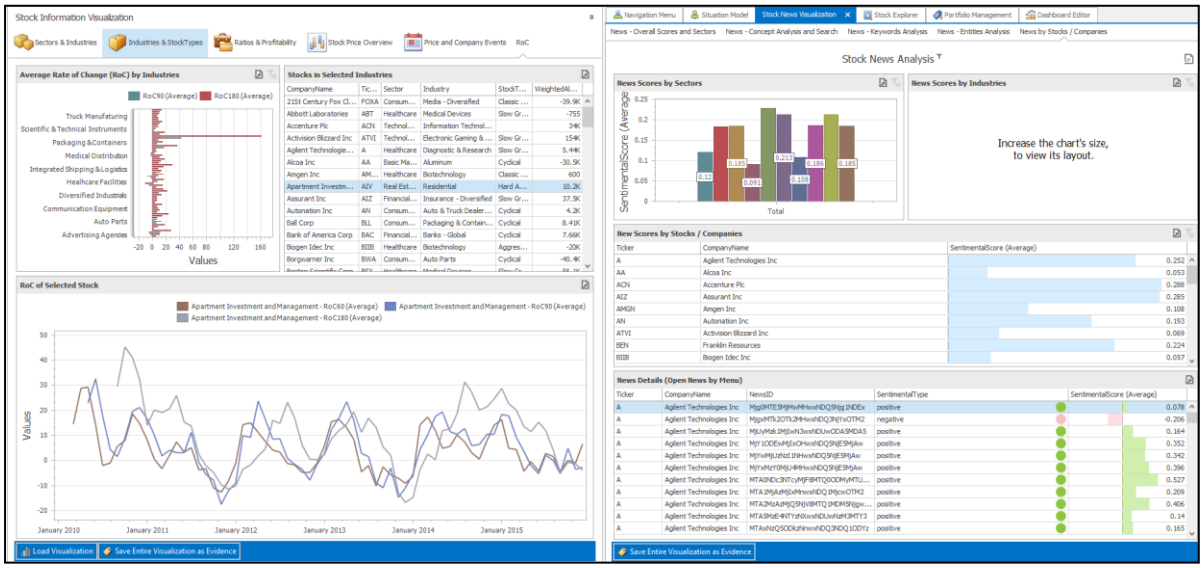
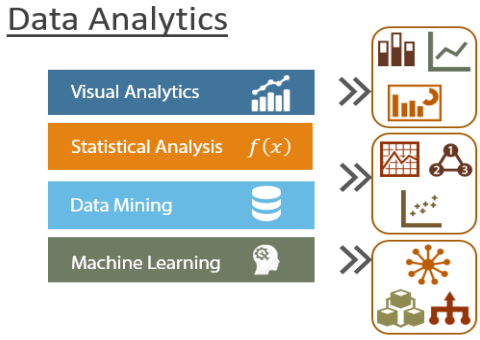


Figure 44. Enquiries are viewed next to each other

---- This Space is Intentionally Left Blank -----

Extra: Why this Study uses a Visual Analytics as the approach for the data exploration phase

There are several data analysis techniques that can be used by data analysts to discover information from the data. As illustrated in the figure below, these techniques include visual analytics, statistical analysis, data mining, and interactive visualization. This study chooses to implement only visual analytics as the data exploration technique.



One of the reasons to choose visual analytics is to control the complexity involved in the data exploration phase of the user study of this study. Visual analytics is an intuitive data exploration technique that can be learnt without significant effort and technical background. This advantage allows the user study to have a larger pool of participants. More importantly, focusing on just visual analytics allows the data exploration performance of the participants to be meaningfully compared.

From the research perspective, the adoption of a visual analytics approach is to represent a data exploration mechanism that is commonly used in the industry. The visual analytics approach that has gaining ever increase attention from both the vendors and the market. These are manifested through the releases of visual analytics products from major vendors such as IBM and SAS in last 2 to 3 years. This is also indicated by the strong growth of visualization-based business intelligence products from QlickView, TIBCO Spotfire, and Tableau.

A similar visual analytics system is included in this prototype to illustrate that the design principles in this study are not developed in isolation from these well-accepted data exploration systems in practice. The design principles can be easily extended to the existing data exploration systems. One of the benefits of positioning our analytic solution as such is that it would seem to be an extension of existing data analytics framework and would have less resistance from the potential users.

5.3.1.4 Intended Effects

The first design initiative – *enabling divergence at data layer* – allows both obviously-related and obscure-related data to be considered from the early stage of the data analytics. This would reduce the chance of the data analysts prematurely narrowing down the scope of data exploration, based on their preconception. It is also speculated that the divergent data encourage the data analysts to search through the full range of data elements that are possibly relevant to the problem situation. Therefore, data divergence enables the data analysts to identify and explore more relevant data elements during their data exploration.

The other design initiative – *enabling divergence at the enquiry layer* – takes the advantage of multi-modal enquiries to help the data analysts to identify important observations. Green et al. (2009) suggested that multi-modal enquiries encourage analysts to derive deeper and more multi-faceted understanding of the data. This allows the users to understand the true nature of the data. With the divergent enquiry, the users will be more effective in discovering implicit relationships between data elements, which in turn will enable them to identify and use more relevant data elements to enhance the quality of their data exploration.

With both these design initiatives 1) enabling divergence at the data layer and 2) enabling divergence at the enquiry layer in place, this study conjectures that the data analysts will be more effective, in terms of searching and filtering, and thus allows them to effectively identify the data elements that are relevant to their problem situation. Therefore, this study conjectures that the data analytics systems that have incorporated the design principle "enabling divergent exploration" will allow the users to perform better in their search & filter activity.

Proposition: The data analytics systems with capability for ***enabling divergent exploration*** will allow users to perform better in the *search & filter* activity.

5.3.2 Enabling Managed Observations

5.3.2.1 Overview of Design Principle

Design Requirement: To support users in capture, manage, and retrieve their observations, including the underlying interpretations.

"Enabling managed observations" is a design principle that stresses the importance of being able to systematically capture, organize, and retrieve the observation and its interpretations. The purpose is to enable these observations to be in a managed state that can facilitate their recall and update, thereby making the observations available to the users for reasoning. This design principle consists of two IS initiatives: 1) enabling observations to be captured; 2) enabling interpretations to be captured. The design principle is alleged to be able to relieve the users' cognitive loads, such as attention span and working memory, allowing them to focus on understanding and interpreting the current observation. The design principle could aid the *perceive* & *interpret* activity.

5.3.2.2 IS Initiative

The first IS initiative aims to support the data analysts to systematically capture their observations together with their enquiry context. Most of the existing works have attempted to support analysts by allowing them to create and organize static annotations about observations made. However, these mechanisms do not extend users' cognitive capacity beyond just having paper and pen to jot down their

observations. The data exploration in complex problems mostly involve enquiries that encode multiple data dimensions. Merely jotting down the observation in static short text would result in the loss of critical details about the observations, thus rendering the observation difficult to recall and less useful for reasoning.

This study proposes that the enquiry context, including 1) the final state of the enquiry and 2) the key attributes of the enquiry, should be captured along with the observation. These two pieces of information enable the users to quickly and accurately recall the observation and thus reduce the need to rerun the enquiry to a minimum. Capturing the enquiry context can be also helpful because observations in complex analytical task often need to be updated to reflect the new knowledge learnt as the user progresses in the data exploration. This design will be helpful for users to recall, reuse or refine the observations.

Design – Supporting the observations and
its enquiry context to be systematically captured

The second IS initiative aims to support the data analysts to capture and make the interpretations of an observation readily available for analysis. Observations are mostly contextualized in the interpretations of the analysts. Researchers also pointed out that in complex problem solving, it is important to go beyond dry data analysis: it often requires the interpretation of the data (Lefebvre, 2004). Interpretations can be viewed as the key properties of an observation. As noted in subsection 4.4.1, the key properties of an observation can be categorized as a data attribute and a derived attribute. A data attribute is based on the value that is explicitly reflected by the enquiry result, whereas a derived attribute is the value derived or inferred, based on the user's subjective judgement. Capturing the observation merely by the enquiry context (e.g. a snapshot data of the chart), which describes the objective side of an observation, results in the loss of the reasoned attributes of the observation, which are often critical for the analysts' reasoning activities in a later phase of the data analytics.

Capturing the interpretation exposes the assumptions and reasoning that the users used in driving the observation, therefore enhancing both the transparency of the reasoning process and the creditability of the observations. Moreover, it facilitates the communication of the information discovered between different analysts. This study proposes to capture the interpretation in structural form and make the interpretation analyzable. The interpretations are stored in computer-recognizable forms which can directly support various computational-based operations, such as criteria-based retrieval, regrouping based on specific attributes, and cluster analysis. It is alleged that this will allow the usefulness of the interpretation to be maximized. The structural form allows these interpretations to be manipulated, analyzed, and systematically presented.

Design: Supporting the interpretation
to be captured in structural form and to be analysis-ready

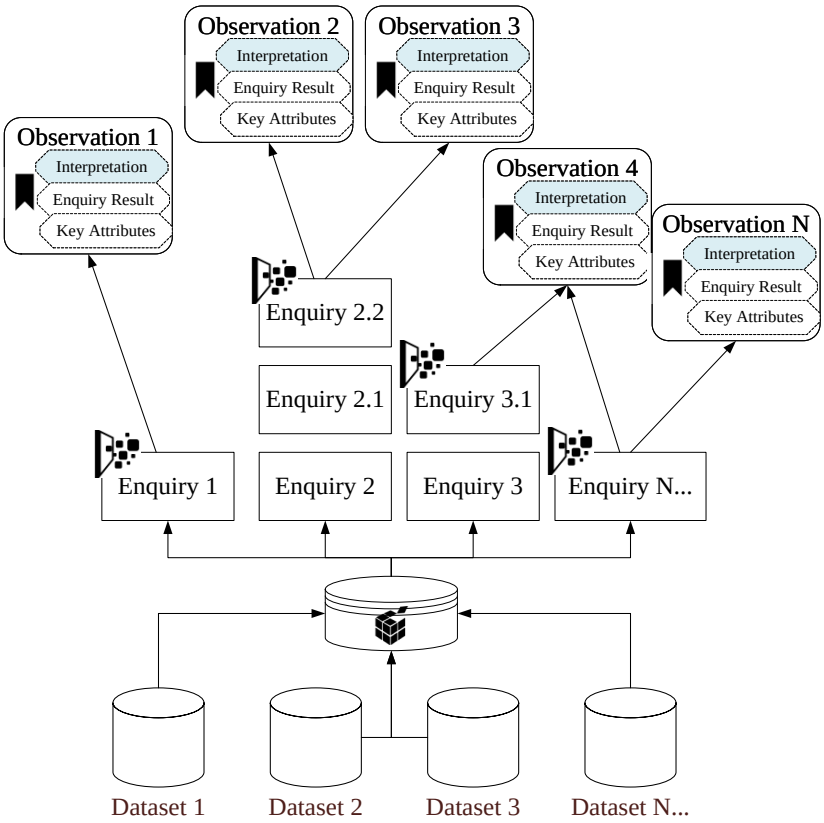


Figure 45. Supporting the observations to be systematically stored, managed, and retrieved

5.3.2.3 Mechanism

The design principle is actualized based on these two IS initiatives. Following the design "***supporting the observations and its enquiry context to be systematically captured***", a design for managing enquiry-aware observations is developed. The design allows the observations to be captured, together with 1) the final state of the enquiry, such as table and charts, 2) the key attributes of the observation, such as the sector of which the stocks belong, 3) underlying data elements being used to compose the state of enquiry (i.e. for chart, this includes the data dimensions at the chart's axes, and data measures for each of the chart's series; for table, this includes the column names and order, and the data in each column).

The following screenshot shows the enquiry is captured into an observation. The data table in Box 1 of *Figure 46* is the result of the enquiry. The users have derived three stock options based on the visualizations above the table. Specifically, the users first investigate the two most promising sectors, healthcare and technology, out of all the sectors (bar charts on top left corner). The users then focus on explore the three industries that have highest *rate of change in stock price* among all the industries from the two sectors. Subsequently, the individual stocks within these three industries are displayed on the line chart. The users have selected the three stock options showing a rising trend in their rate of change

in the last six months. The table at the bottom shows the details of the three stock options, namely Activision Blizzard, Cigna Corp, and Cognizant Technology. Users can save the table as an observation.

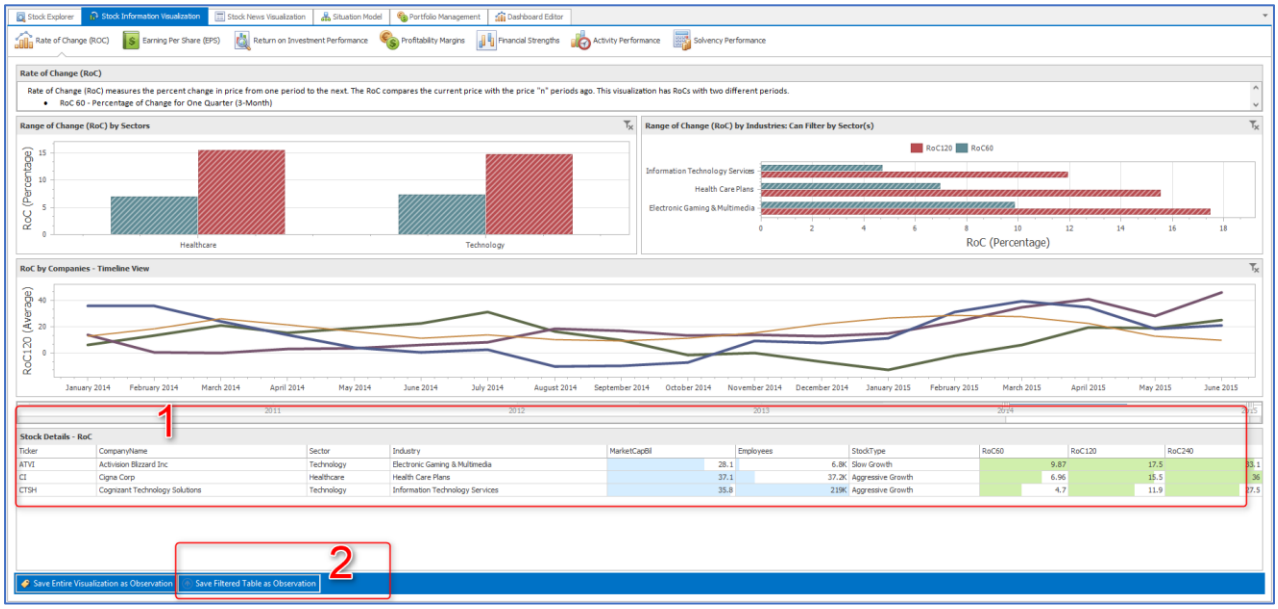


Figure 46. Capturing the state of an enquiry.

Figure 47 shows the interface for saving an observation. The key attributes, such as the selected stock options and the selection criteria, are automatically captured as the attributes of the observation, as shown in Box 4. As proof of the concept, this feature is limited to automatically capturing the stocks filtered (which recorded as Tickers in the interface) and the selection criteria as the attributes of the observation. Box 1 is where the users can enter semantically the title and note for the observation. The *tags* field in Box 2, which allows the users to flexibly define the observation with a user-defined tag, features a hashtag-like function that allow users to create their very own tag. Box 3 shows the automatically captured enquiry state. The enquiry context, key attribute, and interpretation are captured as a single observation.

---- This Space is Intentionally Left Blank -----

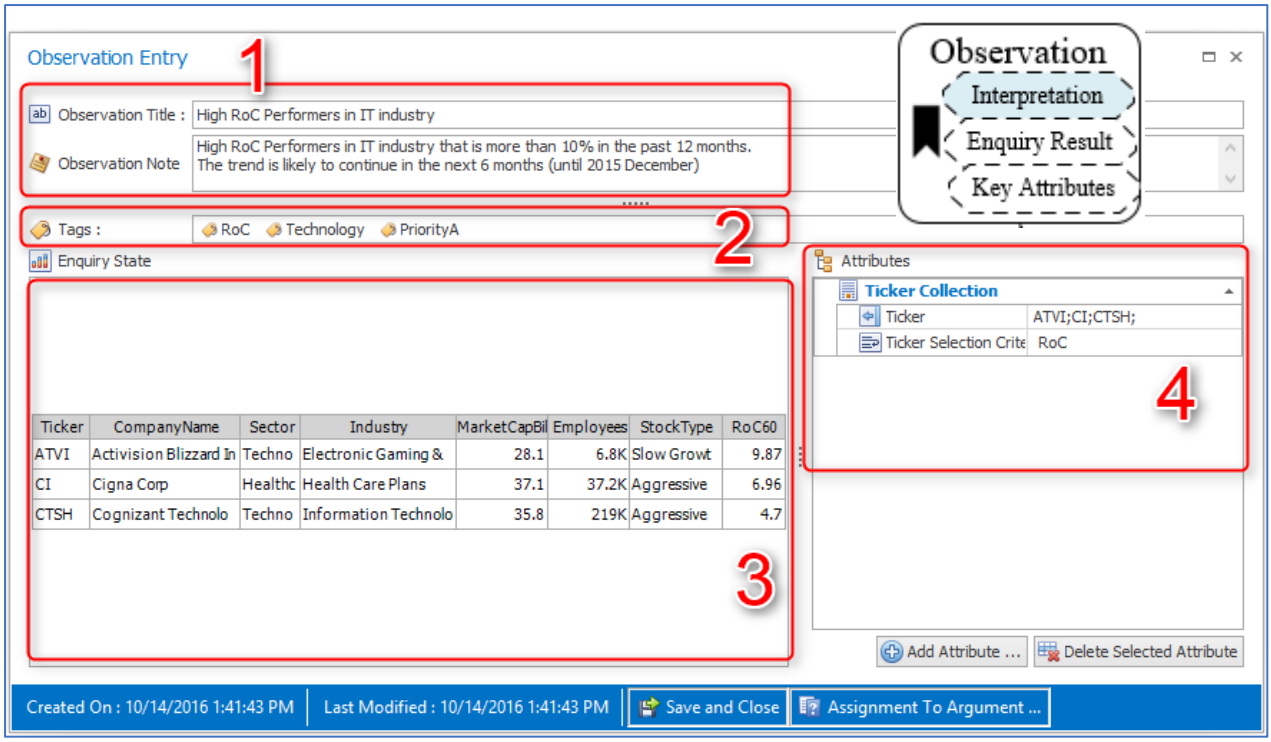


Figure 47. Interface for capturing an observation

The design "***supporting the interpretation to be captured in structural form and be analysis-ready***" allows semantic interpretations from an observation to be stored and incorporated into the data analytics. The specific design suggests that one of the ways to maximize the usefulness of the semantic interpretation is to enable the interpretation to exist in structural form. This study chooses to achieve this structural flexibility by storing the interpretation as a collection of key-value pairs which each describe an aspect of the interpretation. The red box in *Figure 48* shows an example of user-defined attributes that were derived from their interpretation of the enquiry.

---- This Space is Intentionally Left Blank -----

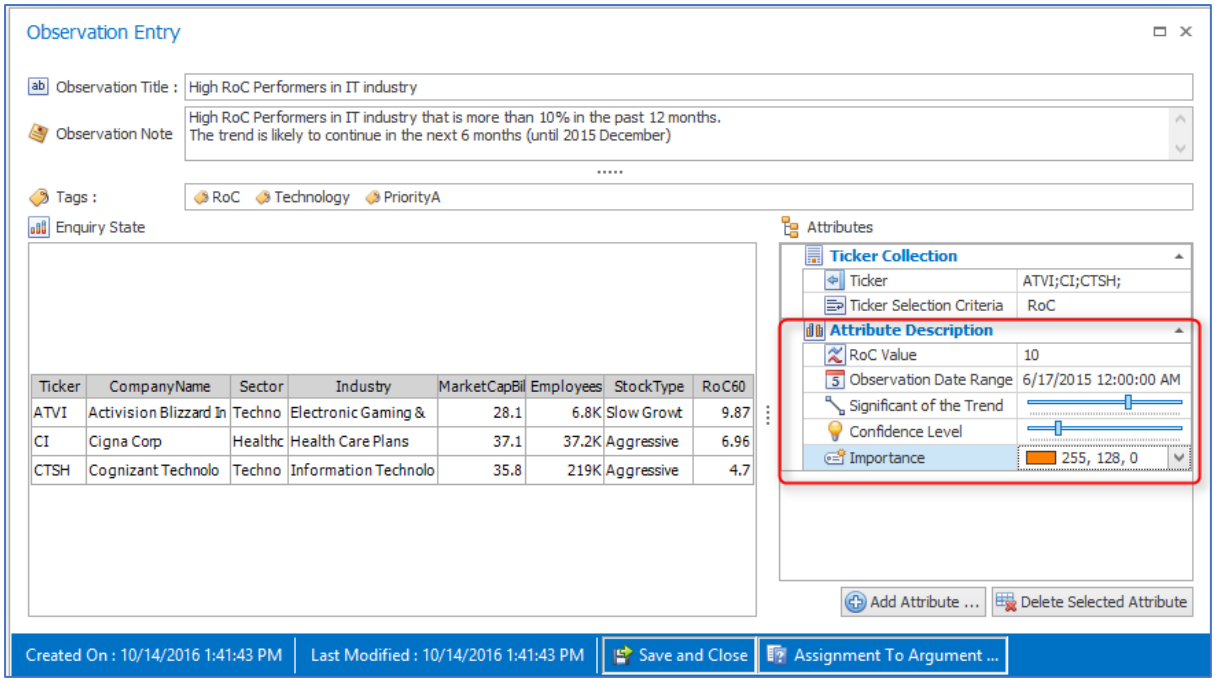


Figure 48. User-defined key attributes of an observation

The design empowers users to freely create a custom list of attributes to describe their interpretations. *Figure 49* shows the interface for managing the attributes. Users can create, delete, and edit the attribute list. They can also use can drag and drop the attributes to rearrange them into the hierarchical structure. This enables the users to use the hierarchical levels of attributes to describe the entity or concept that they want. Once an attribute is created, it is reused in other observations without redefining the attribute. The hierarchical structure will remain when it is inserted into an observation. These user-defined attributes supporting different expressions that could be used to represent the analyst semantical interpretation: a set of common attribute types available to support different type of interpretations. Some of these examples are shown in *Figure 50*.

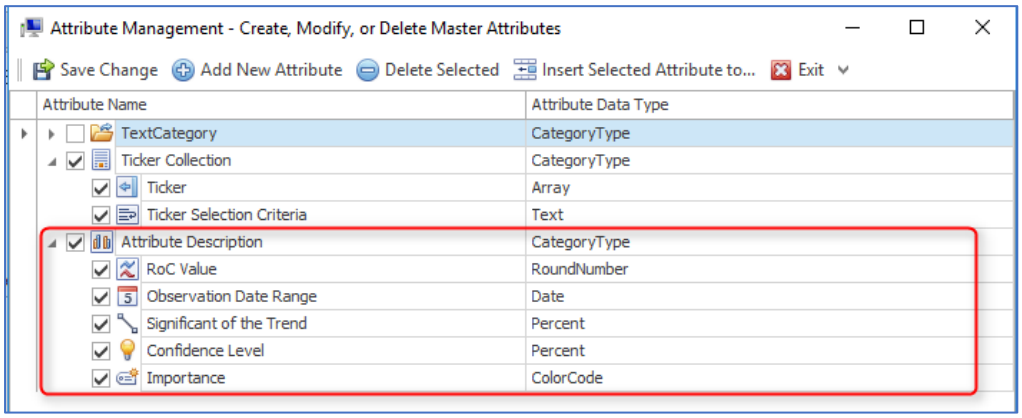
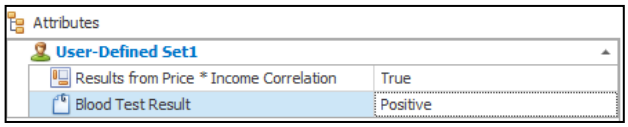
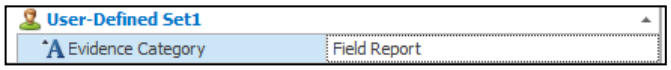
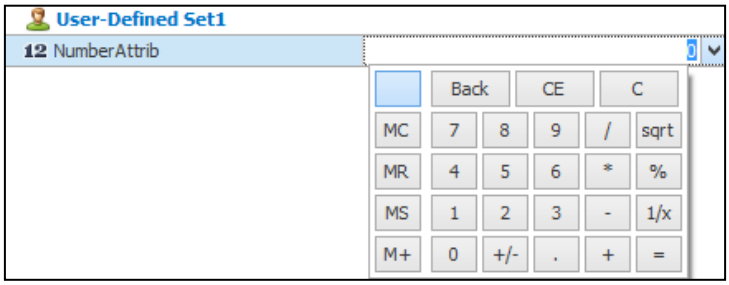
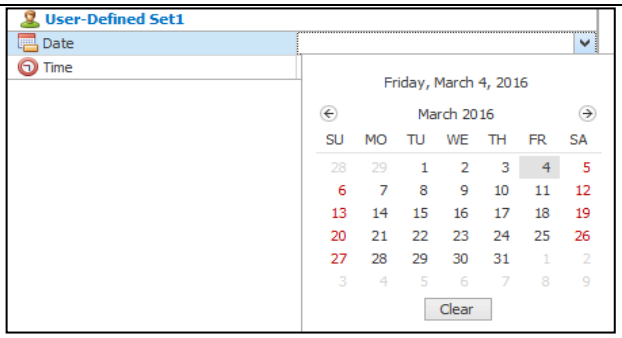


Figure 49. Interface for managing the attributes

Semantic interpretation	Stored Type
▪ Binary reasoning • Logic with two states • Example: Result of statistical test 	Boolean / Text
▪ Categorical • Textual-based characteristics of an object that do not apply to mathematical operations. • Example: ▪ Urgency / Importance ▪ Category of the evidence  	Colour, Text
• Fuzzy Reasoning • Fuzzy characteristics that human normally express as different extents, which often are subjective • Example: Level of Confidence on this particular 	Percentage
▪ Numerical • Characteristics that can be described with number and can be used for mathematics operation • Example: Amount of money 	Number
▪ Date, Time, and Duration • Example: Corresponding timeframe of the price rise	

<table>
<tr>
<td> 

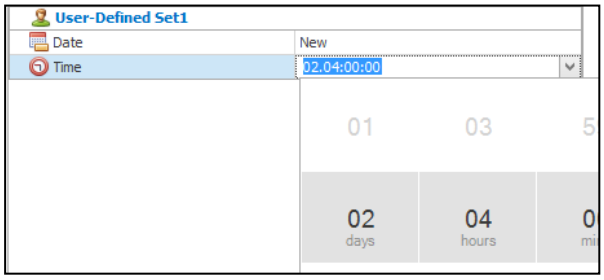
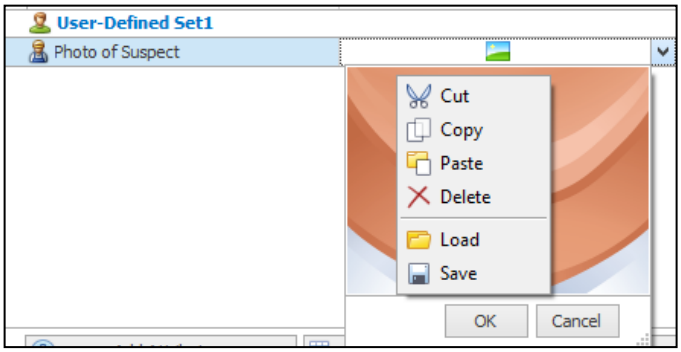

</td>
<td></td>
</tr>
<tr>
<td>▪ Graphic-based
• Graphical representation
• For future development, the semantic meaning of the graphics can be analyzed using IBM Watson’s image recognition and the extracted semantic keywords can be stored as an attribute.
 
</td>
<td>Image</td>
</tr>
</table>

Figure 50. Examples of attribute type can be defined by users

Recall that one of the objectives of managed observation is to allow users to quickly retrieve and recall the observations they have made. Users can generate a list of all observations they have recorded, as shown in *Figure 51*. By selecting any of the observation on the table on the top, the visualization below will show the enquiry state corresponding to the observation. This is alleged to help the users to quickly recall the observation, without the need to open the observation as a full-details view. If necessary, users can double-click on the observation row to open the full-details view.

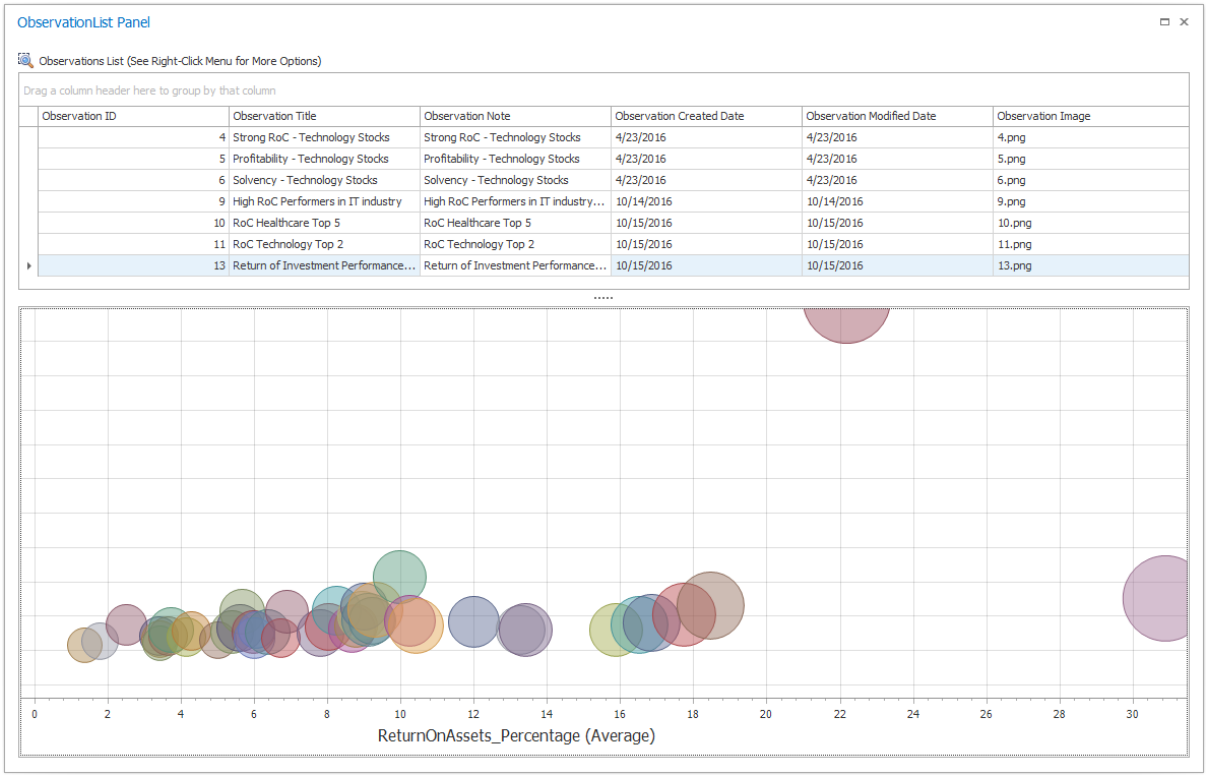


Figure 51. List of observations made

Note that, besides the user-defined attributes, other items of information in the observations, such as observation title, observation note, created date, and modified date, are used for the retrieval of the observations. On the same interface, users can also use search and filter functions to retrieve specific set of observations that meet the search ad filter criteria. These functions are especially useful for complex analytical tasks, which could have large number of observations that accumulated over time. *Figure 52* shows the search and filter functions.

---- This Space is Intentionally Left Blank -----

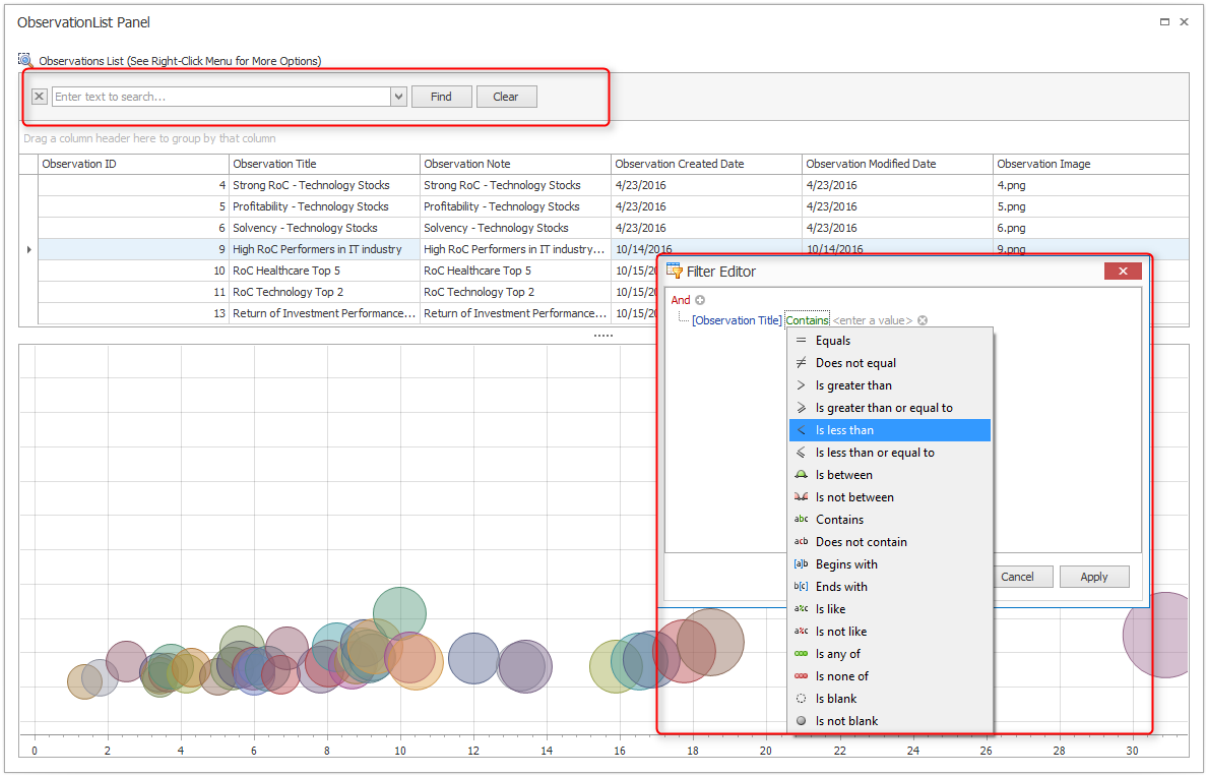


Figure 52. Search and filter functions in the observation list

5.3.2.4 Intended Effects

With the two IS initiatives, 1) supporting the observations and its enquiry context to be systematically captured; 2) to support the interpretations of an observation to be captured in structural form and be analysis-ready, the data analysts would be able to systematically capture, store, and retrieve the observations they made. The designs reduce the cognitive loads required in order to recall the observations when the users need them for analytical reasoning, allowing them to focus on interpreting the data into mental understanding. The design also may help to reduce retention errors, where the observations made are forgotten or cannot be fully recalled. Considering these effects from the design, this study conjectures that those data analytics systems that have the capability for enabling managed observations can help to enhance the user performance in the perceive & interpret activity during the data exploration phase.

Proposition: The data analytics systems with capability for ***enabling managed observations*** will allow users to perform better in the *perceive & interpret* activity.

5.3.3 Enabling Exploration Convergence

5.3.3.1 Overview of Design Principle

Design Requirements: To support users to create a joint summary from their observations.

The design principle "enabling exploration convergence" stresses the importance of being able to have a joint overview of the observations made. The aim is to allow users to effectively learn the overall characteristics of the collective observations, and thus to derive a joint summary from the observations. The joint summary can provide the data analyst with a refined scope of the problem situation which warrants further analysis. This design principle consists of an IS initiative "enabling the visualization and analysis of the meta-observation information", which aims to enable the users to extract and understand the meta-observations information. This study conjectures that this design principle will help data analysts effectively gain implicit and deeper understanding based on a collection of observations.

5.3.3.2 IS Initiative

The goal of the design principle to converge the observations is achieved by the IS initiative that enables the users to visualize and analyze the meta-observation information. This study agrees with the assertion of Mirel and Allmendinger (2004), that the more complex and dynamic the problem situation, the more important it is for the analysts to be aware of what have been found during the data exploration process. Recall that with the features from "managed observation" discussed previously, users can generate a list of observations they have made and open them up to see the details of the observations. This study argues that this feature may be helpful for the users to recall the observations, but is not the most effective way for the users to gain the overall awareness of what they have found so far.

This study further suggests that the user can be better aware of what they have found if the observations can be consolidated and analyzed. This enables the users to make "an observation of the observations", the joint summary across the observations can be more effectively discovered. This is achieved by consolidating and presenting the information across the observation in an interactive dashboard called "converged view". The converged view takes the advantage of interactive visualization and data manipulation to analyze the meta-observation information. This approach will empower the data analysts to slice and dice, aggregate, filter, and visualize the meta-observation information using their own criteria. Therefore, this design initiative proposes a specific design for allowing the users to visualize and analyze the meta-observation information.

Design: Supporting the visualization and analysis
of the meta-observation information

Figure 53 shows the conceptual illustration of the converged view. It is a "view" because a converged view is a summary generated on demand by the data analysts, and it exists temporarily. The converged view is a dynamic existence: every time a new observation is added to the collection, the converged view will change to reflect the new information. Although the converged view itself is dynamic, its state can be permanent stored. If required by the data analysts, the state of the converged view can be captured and stored as a new observation. Notice that not all observations are included in the converged view. Only those observations which meet the data analyst's criteria will be consolidated to provide inputs for the converged view.

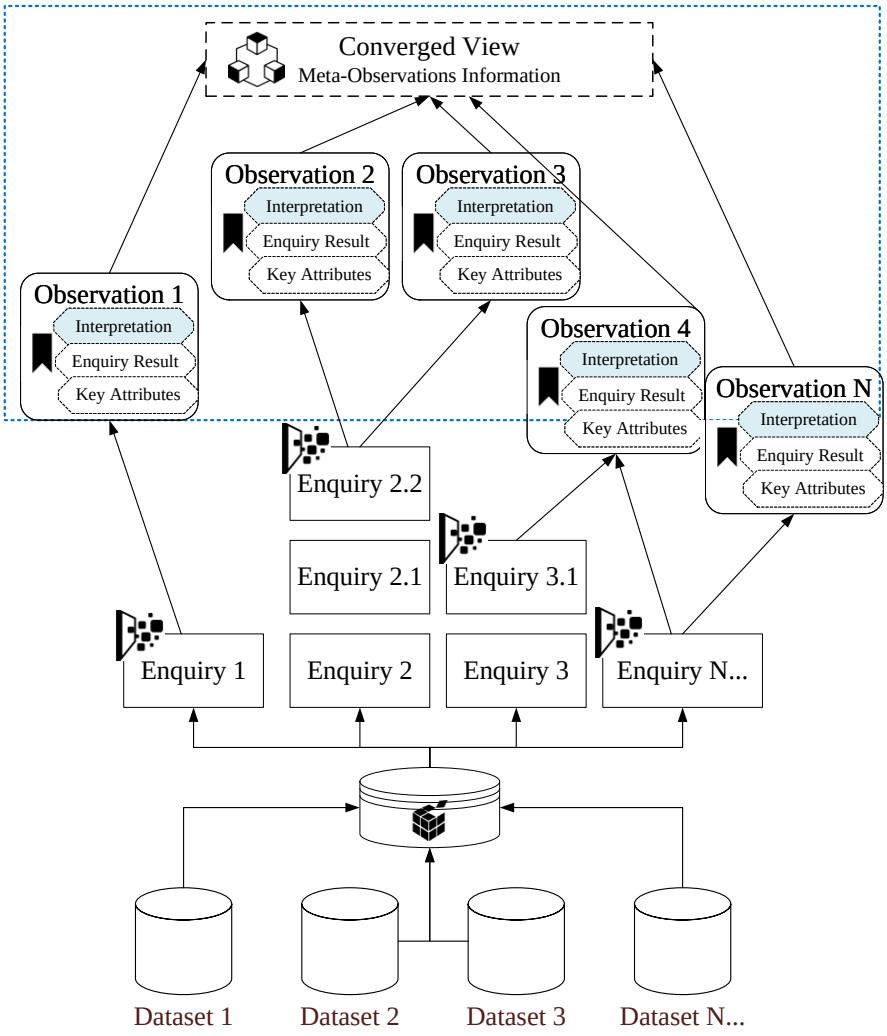


Figure 53. Converging observations

5.3.3.3 Mechanism

Following the IS initiative, a mechanism was developed to actualize the "converged view" concept. This sub-subsection first describes how the converged view can be used by a user, then it explains how the converged view was implemented.

The main objective of the converged view is to allow the users to analyze, visualize, and eventually understand the overall characteristics of the observation collection. As many other analyses, stock market analysis requires the users to investigate the stock options from many different aspects, such as price trend, financial health of the company and sustainability of the price trend. Investigation from each aspect may produce a list of desirable stock options (i.e. an observation), and often the list of desirable stock options is varied by the different aspects the users investigate. Therefore, it is important for the users to have a joint overview of which stock options are the most desirable, from all the aspects. *Figure 54* shows the converged view mechanism in the prototype.

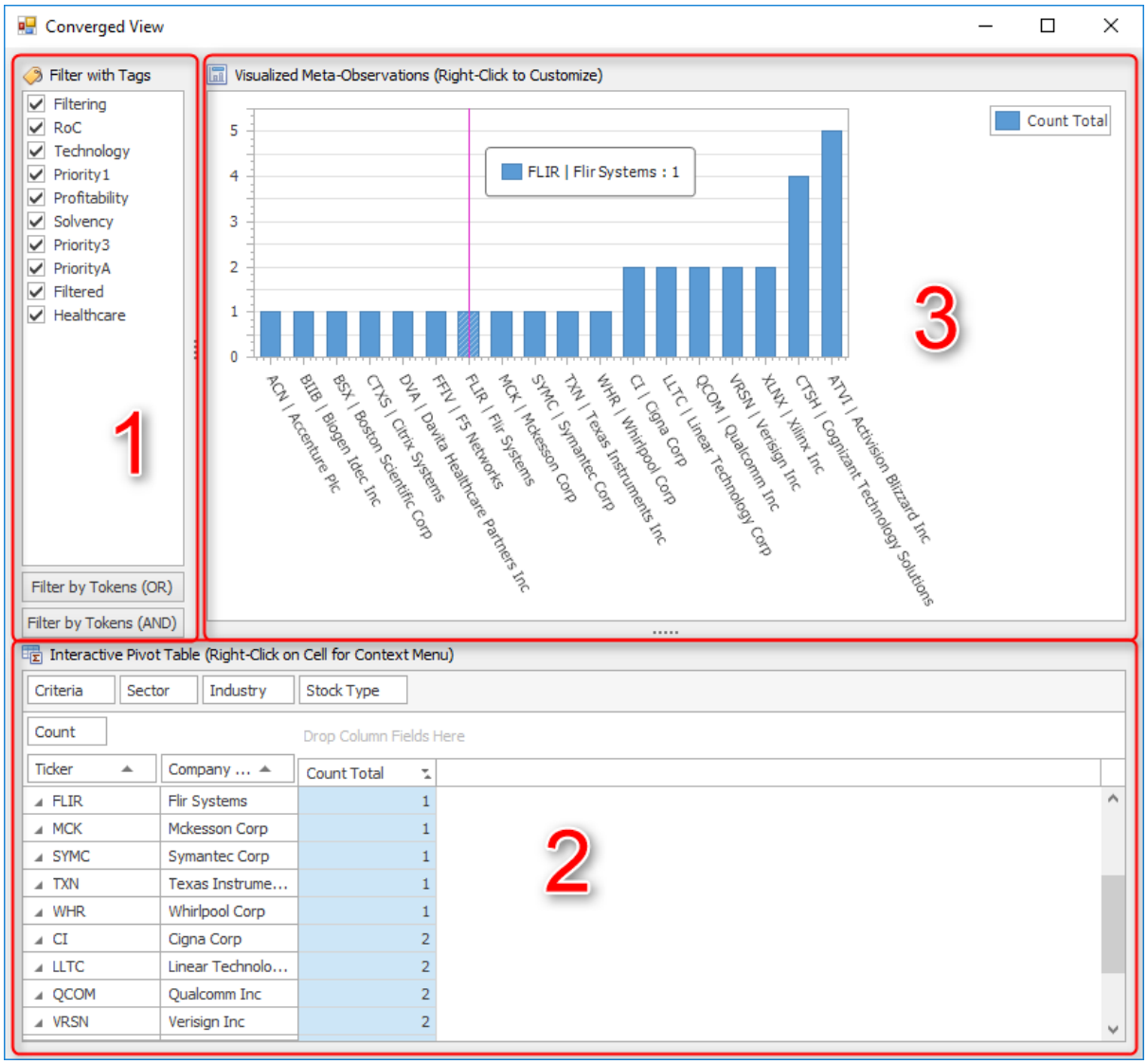


Figure 54. Converged view in overall

The converged view contains three main panels, namely 1) filter with tags, 2) interactive Pivot Table, and 3) Visualized Meta-Observations. The *filter with tag* panel in Box 1 of *Figure 54* can be used to select which observations are to be included as the inputs for the current meta-observation analysis. The tags are the user-defined tag in each observation. Box 2 shows the interactive pivot table,

with which users can filter, sort, and customize the data fields. Then the visualization in Box 3 will update in real time to reflect the users' configuration in the pivot table. The interactive pivot table in converged view is designed in the way that it is similar to pivot table in commonly used spreadsheet application to provide an intuitive way to slice and dice the meta-observation information and reduce the learning curve needed. Users can change or customize the visualization type by clicking on the visualization. Figure 55 shows the visualization configuration wizard.

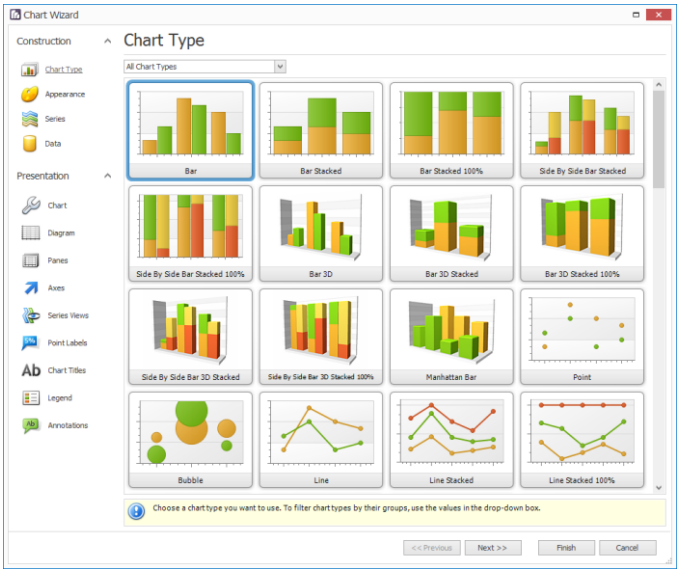


Figure 55. Wizard for customizing the visualization in the converged view

This design initiative is made possible by the previous design principle "enabling observation management". Recall that that design principle allows the observations and their enquiry contexts to be structurally stored. The information sources of the converged view are the observations and their attributes. The enquiry context is made up of the key attributes, which are automatically captured, together with the observation. The automatically captured key-value pairs are obtained from metadata used to construct the visualization or table in the observations. For examples, the metadata of a bar chart created by the users contains the key-value pairs of which data variable was used in the argument axis, which data variables were used in the value axis, and which data series were selected or filtered. As a proof-of-concept feature, the converged view currently analyzes only these automatically-captured key-value pairs. The same concept can also be applied to the key-value pairs created by users. This study suggests that meta-observation analysis can achieve greater potential in supporting data analysts to derive joint summary when both kinds of information are included. *Table 12* shows how all the information captured in the "managed observation" can be used for the meta-observation analysis.

Table 12. The information which can be used for meta-observation analysis

Information in an observation	Description
- Observation Title	- **Text search** Example: Observation that contains "healthcare" in title
- Observation Note	- **Processed text search** Example: Observation that that related to "merger"
- Tags	- **Logical search** Example: Observation that contains tags "grid" AND "observation"
- Creation Date / Last Modified Date	- **Timeline search** Example: Observation that created after 30th June 2015.
- Attribute Name / Value pair	- **Logical search** Example: Observations that have attribute "confidence level" higher than 3 AND "Ticker" contains "BIIB";

Conceptually, converged view is similar a view in database in the way that the result of the query will change to reflect the change in its underlying data. However, the converged view also different from a view in a database in the way that converged views that join multiple existing queries.

5.3.3.4 Intended Effects

With the converged view to visually analyze the inter-observation information, the aggregates, patterns, or outliers across collective observations can be discovered. It could help users to derive joint summary across diverse observations more effectively. This support is especially useful in complex analytical tasks because there can be large number of observations, with each observation containing many more attributes. The converged view makes the large number of observation manageable.

In a user study, D. Gotz and Zhou (2008) found analysts who perform cross-data analysis are more inclined to develop deeper and less obvious findings. This study believes that such positive effects can also be achieved when the data analysts can run cross-observations analyses. This study suggests that the converged view could increase the chances of discovering important and in-depth findings, which are otherwise undiscoverable by looking at each individual observation. Following are other potential reasons why the converged view could improve the user analytical performance.

- More accurate findings – reduce the reliance on impression and gut-feeling to derive the joint summary from multiple observations. The proposed design can take advantage of quantitative approach to present the meta-observation information such as aggregates to help users better perceive the relevant information important to their data analysis, thus allowing them to gain situation awareness level 1 with greater confident and accuracy. This study believes this could reduce the information unavailability errors.

- Reduce time and effort – with the ability to generate inter-observation overview, this support greatly reduces the needs for the data analysts to go through the observations individually. The support provides a scaffolding to retain all the relevant observations and to present the key inter-observations to the data analysts. As the outcome, data analysts can focus their time and effort on reasoning based on the visualized key information.

As such, this study alleges that the design principle "enabling exploration convergence" allows the users to be more effective and efficient in perceiving and understanding the data.

Proposition: The data analytics systems with the capability of ***enabling exploration convergence*** will allow users to perform better in the *perceive & interpret* activity.

In addition, the joint summary that resulted in the converged view provides the data analysts with a better-defined scope of the problem situation. This better-defined scope could smooth the transition from a low-level explorative analysis to a higher-level analysis which has narrower scope but an increasingly deeper analysis.

5.3.4 Enabling Knowledge Creation

5.3.4.1 Overview

> **Requirements:** To support the users to create, manage, and retrieve high-level knowledge based on low-level analytic insights and reasoning.

The design principle "enable knowledge creation" advocates the importance of the ability to create higher-level knowledge from the low-level analytical findings. The objective is to enable the data analysts to use the low-level observations to create high-level knowledge that is meaningful at the problem-solving level. The design principle is achieved through two specific IS initiatives, namely 1) enabling observations integration and 2) enabling synthesis between observations and user reasoning. This study posits that these supports allow users to integrate the low-level observations and synthesize with their knowledge to form high-level understanding about key factors in the problem situation.

5.3.4.2 IS initiative

The design principle contains two IS initiatives that aim to enable users to use low-level data-driven observations to create high-level conceptual factor at the level of which the problem solving operates. The resultant knowledge structure from the two initiatives is computer-recognizable and therefore is ready to be incorporated into the data analytics.

Information integration is common in complex data analytics for two main reasons. Firstly, the complex problem is a huge and high-level problem. The problem often needs to be broken down into multiple interrelated enquiries that run separately (David & Michelle, 2009; Glykas, 2010). Therefore, a single observation provides only a fragmentary answer to the whole problem situation (Zhang et al.,

2008). Therefore, this study suggests that the observations must be integrated to provide the high-level understanding that is meaningful at the problem level. This need is also recognized by researchers from visual analytics field who urge for the need to support the depicting of low-level observations to the higher-level knowledge that emerge from the observations (Thomas & Cook, 2005).

The goal of the first IS initiative "supporting observations integration" is to create a hierarchical knowledge structure which reassembles the way users naturally perceive their analytics problems. The knowledge structure is created by integrating relevant observations into a coherent higher-level understanding. Recall that each of the observations in turn contains the enquiry and interpretation from which the observation is derived. Such a hierarchical knowledge structure is useful for realistically representing the decomposed structure of the big problem. Thus, the users should be supported to systematically integrate the observations they made into such a knowledge structure.

Design: Supporting the creation of high-level knowledge by integrating observations

The objective of the design principle is to enable users to create high-level knowledge that is meaningful at the problem-solving level. The first IS initiative only partially addresses the objective. The nature of the complex problem requires the data analytics to go beyond the fact-driven observation, to require high-level semantic understanding of the problem situation that involving subjective reasoning, such as the analysts' judgement, experience, belief, and grounded intuition. Reasoning is the inevitable logical inferences in the process of creating the high-level knowledge. However, reasoning can be 1) highly subjective as it is largely informed by the data analyst's tacit knowledge, 2) very messy and potentially easily forgotten even by the analysts themselves. Together, the reasoning and the integrated observations form the complete high-level knowledge – an argument.

This study suggests that the reasoning used to support an argument should be captured. This is because when the argumentation process is complex, it is important to externalize the reasoning and assumptions (Yedendra B. Shrinivasan & Wijk, 2008). The validity of the data analytics hinges heavily on the reasoning that underpins the argumentation. Externalizing the reasoning allows the argument to be easily traced back to its source. Moreover, the reasoning is a key information to help other data analysts to understand and collaborate in the data analysis. This IS initiative is an endeavor toward promoting the elucidation of the reasoning used in data analytics. This "white-box reasoning" approach aims to provide the transparent, traceable, and learnable justifications for the knowledge created. Amar and Stasko (2004) have argued that this approach is a solution to the "rationale gaps" which are part of the obstacle for data analysts to conduct higher level analytical tasks.

The second IS initiative is to enable the users to formalize and capture their reasoning use. As illustrate in *Figure 56*, there will be a set of reasoning per argument, regardless of the number of

observations used to support the argument. As indicate by argument 3 in the figure, an argument can exist without observations. Argument without observations are the result of argumentation entirely based on subjective information. This flexibility allows the data analysts to incorporate subjective knowledge beyond those available datasets, to include informal information found online, previous experience, company policy, or even result of decision collectively made.

Design: Supporting the reasoning used to form a higher-level knowledge to be structurally captured, stored, and retrieved.

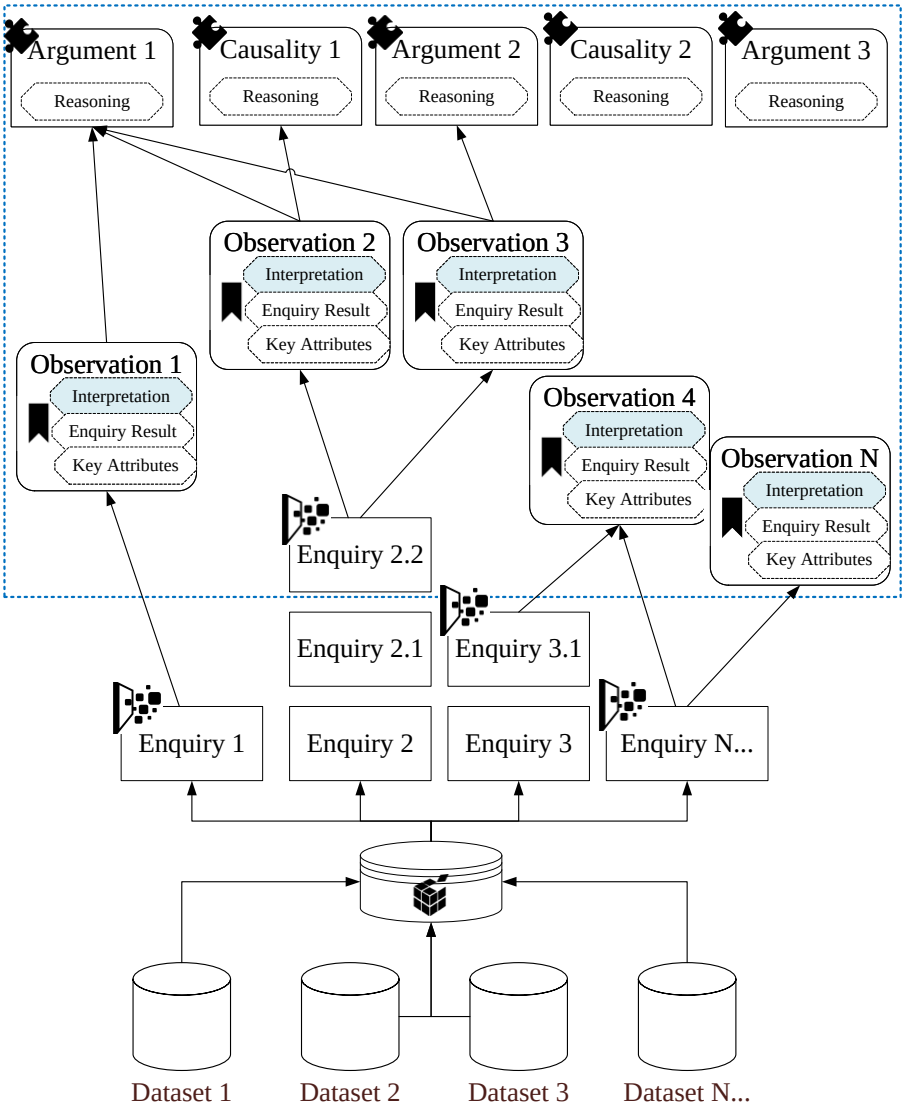


Figure 56. Enabling knowledge creation by integrating observations and synthesizing reasoning

5.3.4.3 Mechanism

At the actualization of the IS initiatives, a mechanism is implemented to allow the data analysts to carry out the observation integration and synthesis process required to create an argument. The user interface

to access this mechanism is called the “argument dialog” in the prototype. The user interface contains three main tabs: 1) basic information, 2) observations attached, and 3) editor.

The observation integration can be achieved using the “observation attached” tab. Users can associate an argument with zero to multiple observations under the “observation attached” tab and can invoke the observation selection dialog being shown as the overlying window. The observation list dialog allows the data analysts to quickly retrieve previously made observations to be associated with an argument. Specifically, the data analysts can retrieve the observation based on three features: quick preview, quick search, and detailed view.

- Quick preview – when an observation is selected, the quick preview will instantly display the “state of enquiry” to allow data analysts to choose the right observations without drilling down to the details of the observation.
- Quick search – will allow effective retrieval of relevant observations based on the search terms. This feature is helpful when data analysts need to retrieve observations from a large collection of observations. Observations can be searched based on keywords in title, descriptions, created date.
- Detailed view – a double click on an observation will open a full detailed view of the particular observation in a separate dialog. Data analysts can make direct changes in the detailed view and these changes will be reflected in all arguments that are associated with the observation.

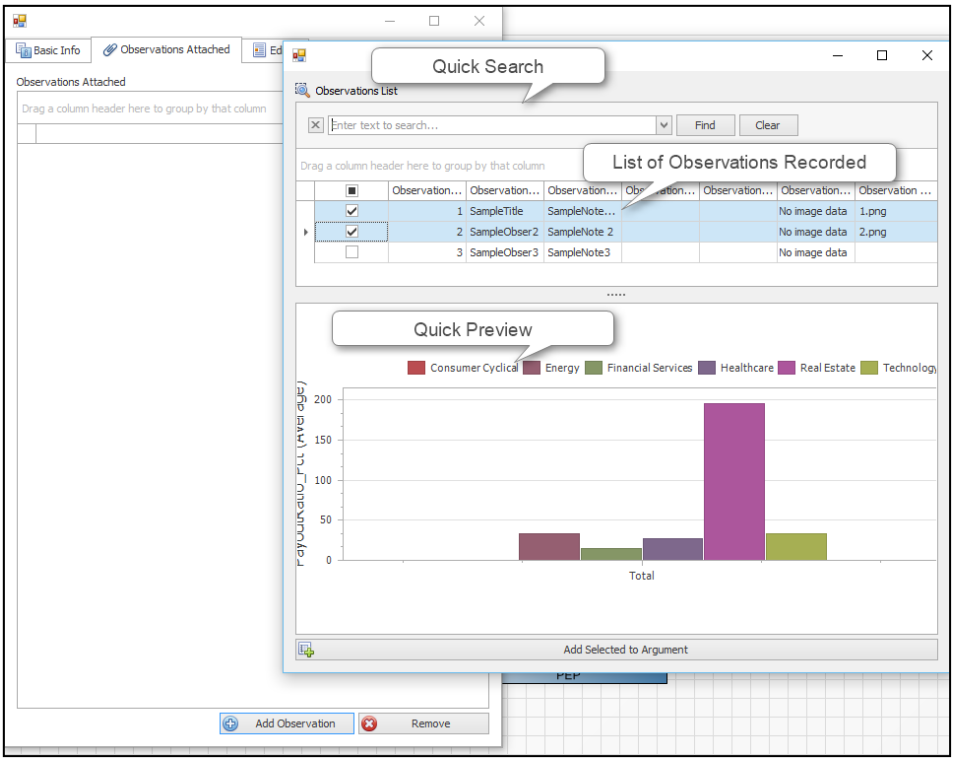


Figure 57. Selecting observations to be associated with an argument

Figure 57 shows that the two selected observations have been attached to an argument. This means that this argument is supported by two observations. This feature allows the data analysts to create a conceptual-level factor based on the integration of multiple observations. It also allows an observation to be reused in multiple arguments. However, caution should be taken when the users notice they heavily use a single observation to support many arguments in their analysis. In such a case, the observation becomes a critical point which the analysis is heavily hinged on. The observation should be scrutinized to ensure all the information and interpretations are correct. During the situation modeling phase which will be presented in the next subsection, arguments are presented as the nodes in a Bayesian network model. In this current prototype, the information stored in an argument will be automatically reasoned in order to establish the parameters of the node in the network model. Users need to manually set the parameters of the nodes by inferring the information in the argument.

The creation of the argumetn would not be complete without the support to allow the users's subjective reasoning to be synthesized with the fact-driven observations. The reasoning are important as they contain the logical inferences, external information, and premises that justify how the individual observations are being synthesized into a coherrent argument. The prototype has two ways to capture and store the analysts' reasoning: 1) attribute name-value pairs on the "Basic Info" tab and 2) text composer on the "editor" tab.

The attribute name-value pairs made up a mechanism similar to the one introduced in observation management. It allows data analysts to flexibly describe the characteristics of the argument. As shown in *Figure 58*, the argument derived from the two observations is "ATVI has had a strong position since 2014 June"; the argument consists of three attributes, including Ticker, Trend, and Market Poll Opinion. Trend is a subjective inference that the analyst made based on the informaiton in both observations. Market Poll Opinion is a example of external information (often informal information) which the analyts concludes after reading online investment articles or discussing with fellow analysts. Other information such as confidence level of the argument can also be entered as an attribute.

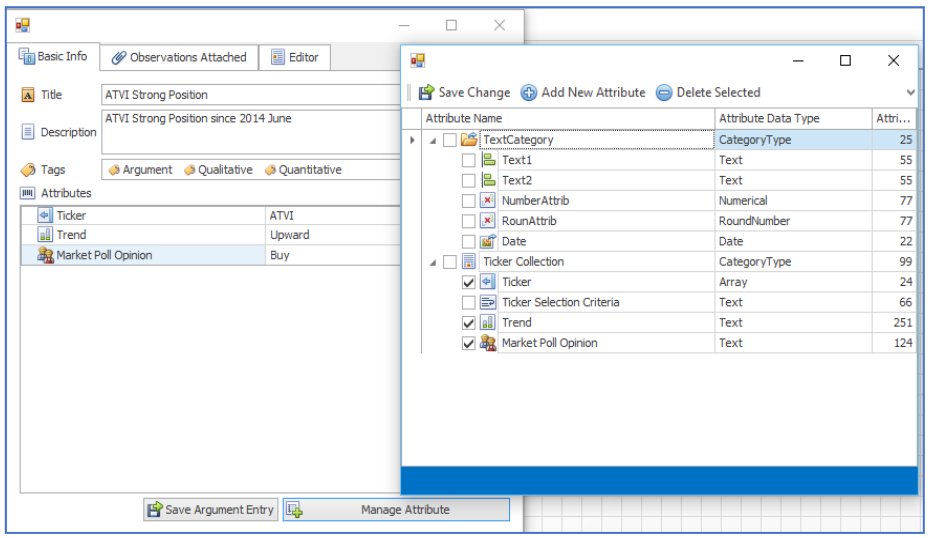


Figure 58. Enabling reasoning to be captured as the attributes of argument

The second way to capture the reasoning is through the text composer. While the attribute provides a clean and structural way to capture the key elements in the reasoning, the attribute name-value pair mechanism might not be sufficient to capture the richness and flow of a coherent story. The text composer allows the data analysts to fully narrate the logical inferences, assumptions, and information sources using the open style that most people are familiar with. Data analysts can take the advantage of various representations such as table, picture, text and diagram to articulate the reasoning. Recall of the argument can also be formed without referred to any observation. In a complex problem, it is common that the datasets do not contain all the information needed, so the data analysts would have to seek for information from various external sources. These sources could be webpages, offline report, and outcomes of a meeting. This information can be inserted into the composer to incorporate it into the data analysis.

Once data analysts have entered information into the composer, they can run a text analysis to automatically capture key elements in the content, as shown in *Figure 59*. This is useful for the data analysts because, although the composer is valuable to fully narrate the reasoning or external information, its length and unstructured content make it difficult to be quickly understood by the data analysts for analytical reasoning. The text analysis provides the ability to extract the key elements (see *Table 13*) from the composer's contents into a structured form. These generated key elements will be stored together with the argument. Because they are stored in structured form, the elements can be used in analysis with the other attributes name-pair value.

Table 13. Information can be extracted from the texts

Key Elements can be Extracted	Description
• Keywords	• Identify the keyword used in the content.

• Entities	• Identify people, organizations, location, cities, and other entities within the composer's content.
• Concepts	• Identify the overall concepts of the content.
• Classification	• Categorize the content into a higher-level topic category such as business, finance, technology, or healthcare.

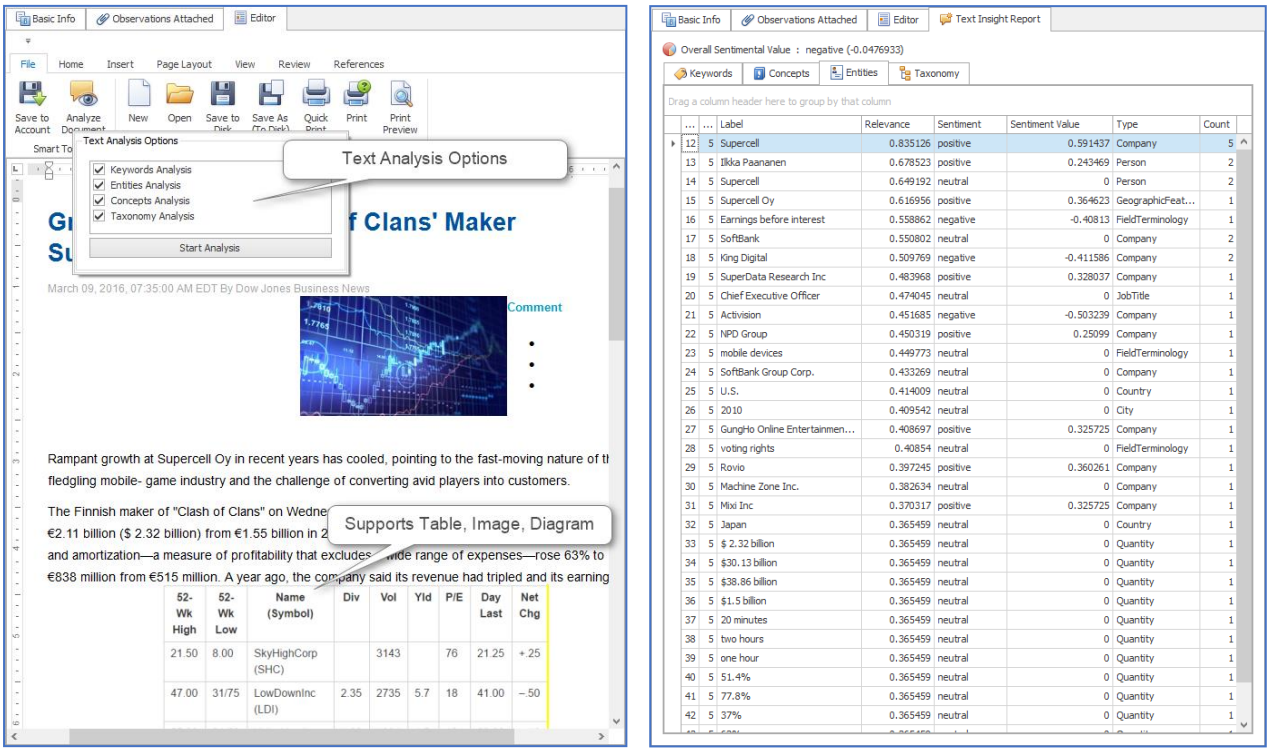


Figure 59. Text analysis extracting user's reasoning into structural information

As noted, the argument derived from observations and subjective evidence is synergic in nature: the whole of the argument itself is greater than the sum of its lower-level elements. The analysts derive some contents, such as the reasoning that they analysts used to derive such argument, from the combination of observations and subjective evidence. This is the information that encapsulates the domain knowledge of the analyst, which is often hard to capture in a rigid structure. To allow the full extent of freedom for the analysts to express their reasoning behind the argument, conventional method is to choose to store this as a media rich format in which analysts can use hyperlink, image, drawing, and video for the expression.

However, this results in the loss of its capability to be actively included in the data analytics. The design in this study proposes the processing of such rich content to produce structural information, and even to extract the hidden information from such rich information: information that is important to the analysis. This can be achieved mainly by semantic analysis and image processing. As for this prototype, it only process and present the structured information to the users as the result, the users have to decide

and to enter the relevant information as the key-value attributes of the argument. For the future development, the systems should be designed so that the produced structural information can be easily selected to enter as attributes of the argument.

5.3.4.4 Intended Effects

The designs from the design principle "enabling knowledge creation" simulate how expert analysts work. As stated by van Merriënboer and Sluijsmans (2009), experts have mental schemata that allows them to treat a set of interrelated elements as one single element, allowing them to cope with higher element interactivity. The designs achieve this in two ways.

Firstly, the design allows the users to integrate the low-level technical to go beyond information recall, in order to create a higher-level conceptual understanding that is meaningful at the problem-solving level. Conventionally the "new knowledge" is created in the form of static text annotation which the users can enter, going manually through the observations one by one. The proposed design allows the users to quickly and accurately retrieve the relevant observations to create the new knowledge. This support is especially useful for the users to cope with the large number of observations common in a complex analytics problem.

Secondly, the design allows the users to synthesize their reasoning with the fact-driven observations. The reasoning can be quickly captured in a structural form to facilitate clear understanding of the assumptions and judgment used to support the argumentation. The design makes the reasoning that drives the argument explicit, thus making it easy to be assessed by the data analysts or others for identifying any weak spot which could invalidate the argumentation.

Proposition: The data analytics systems with capability for ***enabling knowledge creation*** will allow users to perform better in the *integrate & synthesize* activity.

5.3.5 Enabling Assisted Situation Modeling

5.3.5.1 Overview

- **Design Requirement**: To support the users in identifying a preliminary core structure of the situation model
- **Design Requirement**: To support both quantitative and qualitative approaches to situation modelling
- **Design Requirement**: To support the users in constructing interactive, dynamic, and computation-friendly situation models.

The design principle "enabling assisted situation modelling" stresses the importance of scaffolding the users' situation modelling process, with the objective of reducing the process complexity while enhancing the quality of the resultant situation model. The goal of the design principle is to enable users to gain a holistic understanding of the problem situation. In conjunction with this goal, researchers have also claimed that the design that aids users in a form that facilitates understanding of problems, should be included as one of the primary objectives of providing information systems support (Thomas & Cook, 2005). The design principle consists of three specific IS initiatives, namely 1) enabling the selection of the core structure of a situation model, 2) enabling both quantitative and qualitative approaches to situation modelling, and 3) enabling a dynamic situation model that can facilitate rich interactions between the analysts and the model. This study conjectures that with these supports in place, the quality of the situation model can be enhanced and the complexity of the modelling process can be reduced. As a result, data analysts are more likely to be more effective, in terms of the *connect & build* activity.

5.3.5.2 Design Initiative

The goal of the design principle is achieved through three IS initiatives: 1) supporting the users to identify a preliminary core structure of the situation model, 2) supporting both quantitative and qualitative approaches to situation modeling, and 3) supporting the users in constructing interactive, dynamic, and computation-friendly situation models.

The IS initiative "***enabling the selection of the core structure of the situation model***" aims to support the data analysts in identifying the core structure of the situation model, even before the users start the modeling process. Rudolph (2003) found that participants who jumped to an early conclusion and fixated on it showed the worst performance. Surprisingly, the participants who kept an open mind and refused to speculate were just mediocre, and not the best. The best participants were the ones who jumped to an early speculation but then deliberately tested it. Applying this understanding in complex problem solving, this study infers that the users who have a preliminary frame of how things work in the problem situation and then deliberately find evidence to improve and test the frame will have better analytical performance. In the context of this study, the preliminary frame refers to the core structure of the situation model, which allows the arguments and causalities to fit into it to form the big picture of the problem situation. This argument is aligned with the common notion of sensemaking, where the sensemaking is a process of collecting data with a temporary frame, and then fitting the frame around the available data. The two loops are intrinsically connected (Weick, 1995).

However, not everyone has the knowledge or experience to generate the preliminary core structure of the situation model. Experienced domain experts are able to construct more complete and accurate situation models because they have richer mental schemata that allows them to understand a wider range

of causal connections that govern how things work in the problem situation. Although experts' mental schemata are hardly accessible, fortunately the situation models that the experts created can be formalized and transferred. Academic and industrial research has developed conceptual frameworks for various domains, based on the inputs of domain experts (Demirer, Mau, & Shenoy, 2006; Shenoy & Shenoy, 2000). Some of these frameworks have been replicated by other researchers to rigorously test for their validity and robustness under different cases. This study suggests that these conceptual frameworks can be used to inform data analysts about the core structure of the situation model. Specifically, the conceptual frameworks from experts or research can be used to outline the structure of major components in the problem situation and how they are related to each other. Then, novice users can use this core structure as the start-up point for building their situation model. The analysts can build their problem-specific situation modeling by expanding or modifying the base model. This support will allow the data analysts to take advantage of a well-tested conceptual framework to confidently produce their situation model. *Figure 60* shows the conceptual illustration of the core structure.

Design: Supporting the use established conceptual framework as the core structure of the situation model.

---- This space is intentionally left blank ----

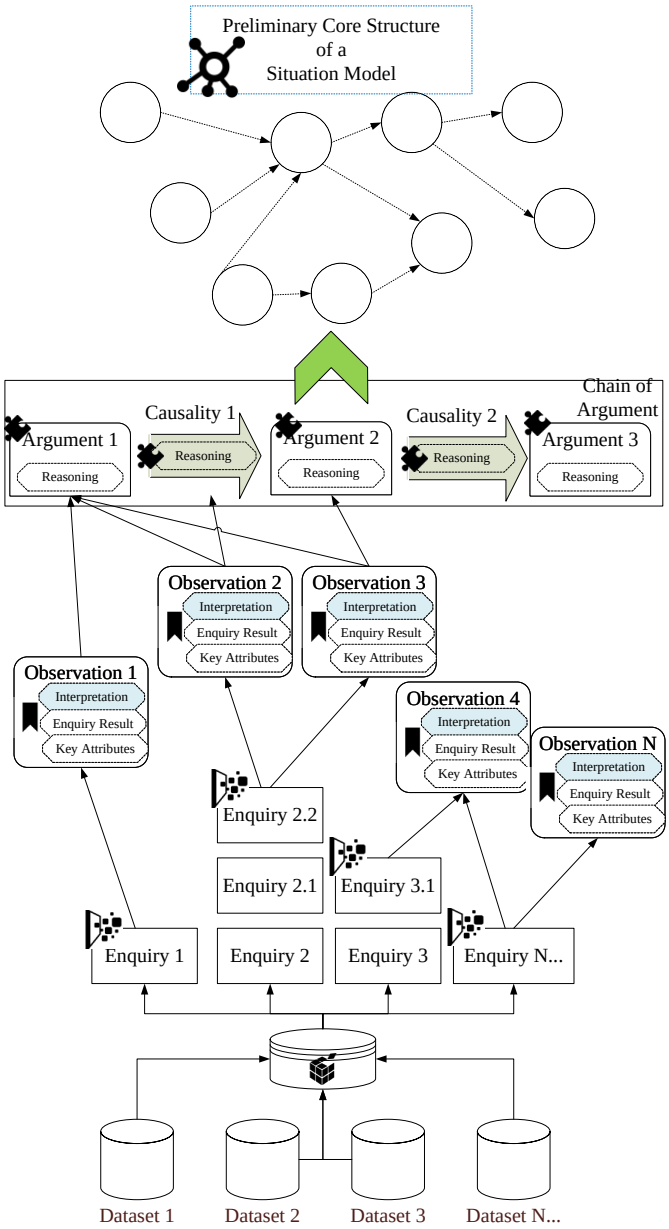


Figure 60. Supporting the users to identify the core structure of a situation model

The design initiative "***enabling both quantitative and qualitative approaches to situation modelling***" is important for complex analytics problems. Due to uncertain and incomplete information in complex problems, data analysts often need to integrate quantitative and qualitative information in a complementary manner to construct a complete situation model. However, most of the systems that support situation modelling rely almost entirely on the analysts' subjective intuition and judgment to establish the arguments and their relationships. Sole reliance on the users is prone to biases and errors. Another problem with such an approach is that it forfeits the power of the quantitative information to in establish and validate the situation model. Therefore, there is a need to support both quantitative and qualitative approaches to situation modelling.

This study proposes to tightly couple both human and machine approaches for building the situation model. Human analysts have the context-rich reasoning to allow them to identify the relationships that are semantically important, whereas computation aids can use statistical inference to rigorously assess the strength and confidence of such relationships from data available. The process will be initialized by human analysts to identify the relationships; then the machine become responsible for providing empirical assessment on the relationships. This empirical assessment provides suggestive feedback to the analysts; but the decision whether to retain or remove the relationships is based on the analysts' judgment on the empirical assessment of the relationships.

Moreover, the design also allows the role of the users to move from being a passive "approver", accepting or rejecting the suggestive feedback from the machine to being an active "builder" in the situation modeling process. The design enables data analysts to build situation models using both quantitative and qualitative information. In a situation model, quantitative information from the datasets can modeled as *data-driven factors*, whereas the qualitative information includes the argument and association that users derived from observations and reasoning. Although the argument and association are fact-driven information derived from observation, they are considered as qualitative information in the design because they are the results of the users applying their reasoning and interpretation to extract information from the observations. Both these qualitative and quantitative information can be used to build a situation model. Where the data-driven factor is unavailable, the argument-driven variable can be used to complement the information in the situation model. This approach allows the users to actively use their domain knowledge as the building blocks for the situation model. Moreover, the data-driven factors are constantly connected to the data source. This makes it possible for the situation model to be automatically updated as new data is streamed to the data source.

Design: Supporting the situation modeling that can have
both data-driven factors and argument-driven factors to build a situation model.

---- This Space is Intentionally Left Blank -----

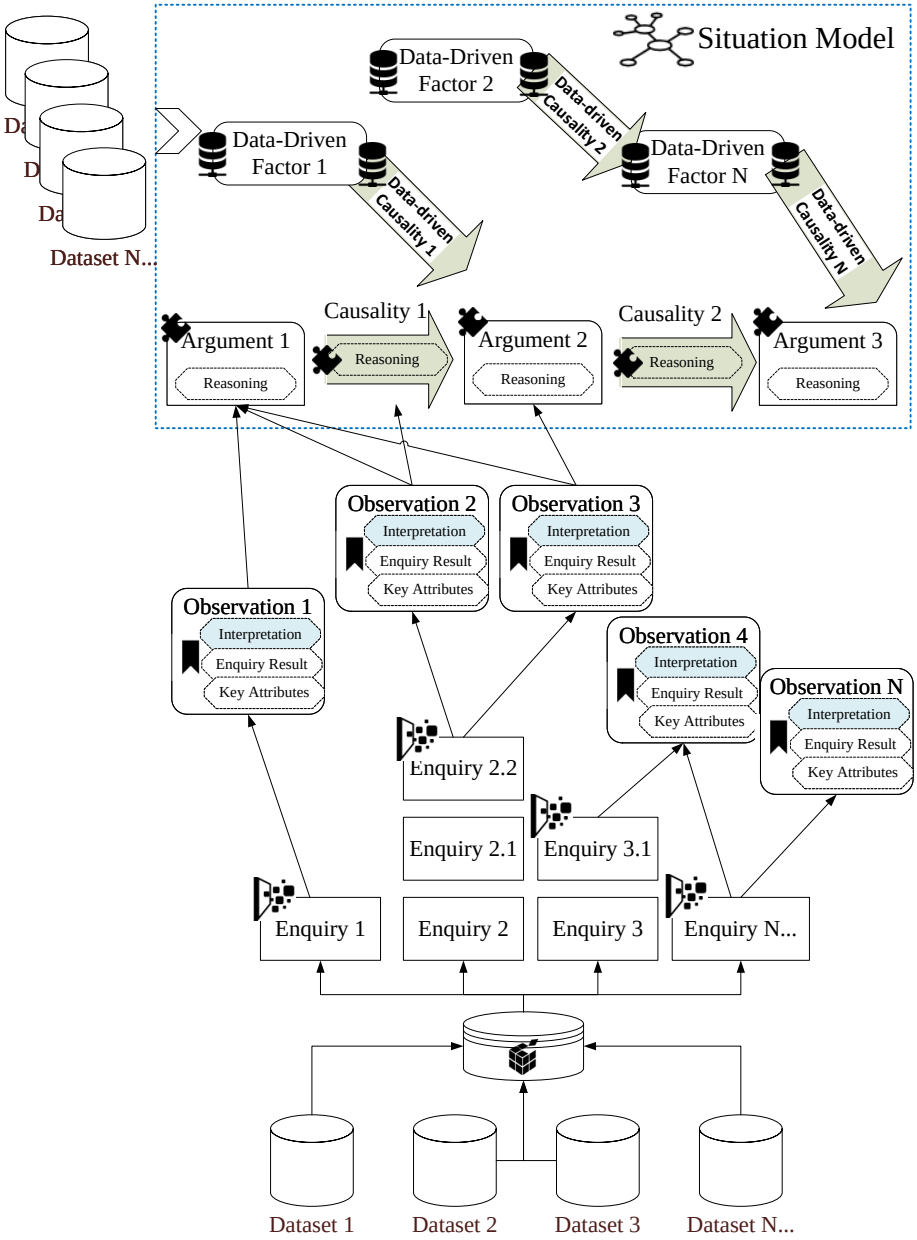


Figure 61. Supporting the situation modeling with both quantitative and qualitative approaches

The third IS initiative "***enabling the dynamic situation model***" aims to produce a modeling technique that allows the analysts to intuitively articulate their subjective view of the real-world problem situation into a dynamic situation model which will allow them to comprehend the holistic view of all the analytics findings at the level of significance for informing the decision-making action. To allow the analysts to intuitively articulate their view, this study proposes the use of "visual modeling" as the input technique, as it allows the analysts to specify their model through an interactive visual interface. For enable the dynamicity of the model, this design in this study adopts computer reasoning techniques combined with statistics analysis methods. Specifically, this study uses Bayesian network modeling and multiple regression modeling as the underlying engine for the dynamic situation model. Bayesian network modeling is a graphical probabilistic modeling technique widely used for knowledge

representation and reasoning under uncertainty. Using this technique allows the visual modeling developed in this study to have the following benefits:

- ➔ **Visual modeling** – the design enables data analysts to specify their model through graphical interactions such as drag-and-drop, selections, and slide bar. This provides an intuitive and quick way to specify the models and does not require the data analytics to learn new modeling syntax. Visual thinking is aligned with the way data analysts build the mental model naturally.

- ➔ **Hidden complexity** – the interaction is used to specify the underlying technical operations without requiring of the technical knowledge. Specifically, the interaction between the users and the modeling tool translates the parameters of the model into mathematical equations.

- ➔ **Integrating information** – the modeling technique allows users to integrate information from various sources within a single model. The users can combine their domain knowledge with the quantitative data.

- ➔ **Robust to Uncertainty** – the modeling technique allows uncertainties in each factor in the situation model to translate into uncertain in the final prediction. More importantly, it allows assumptions about cause and effect in the situation model. It is valuable to represent the situation model with uncertainty, unpredictability, and imprecision.

- ➔ **Prediction Capable** – the situation model can represent holistic of the system by factoring in causal relationships into a coherent predictive-capable model.

Design: Supporting visual modeling technique for specifying the situation model that is interactive, dynamic, and computation-supportable.

5.3.5.3 Mechanism

As the actualization of the design "***Supporting the use of established conceptual framework as the core structure of the situation model***", a mechanism is implemented to allow the data analysts to choose from a list of "model schemas" when they want to create a situation model. These model schemas are established frameworks for the analysts to use as the starting point. The source of the established frameworks can be industrial-wide practice, knowledge solicit from experts, or well-validated research findings. These established frameworks should contain only the core skeleton structure so that it minimizes the chance of imposing rigid thinking on the analysts. Two well-accepted conceptual models of the portfolio return are made available for the users in the prototype. *Figure 62* shows the interface

in which the users can start creating a situation model by selecting the button in Box 1; then a pop-up dialog in Box 2 will show up and display a list of model schemas. Users can choose any of the model schemas as their "template" to start the situation modeling.

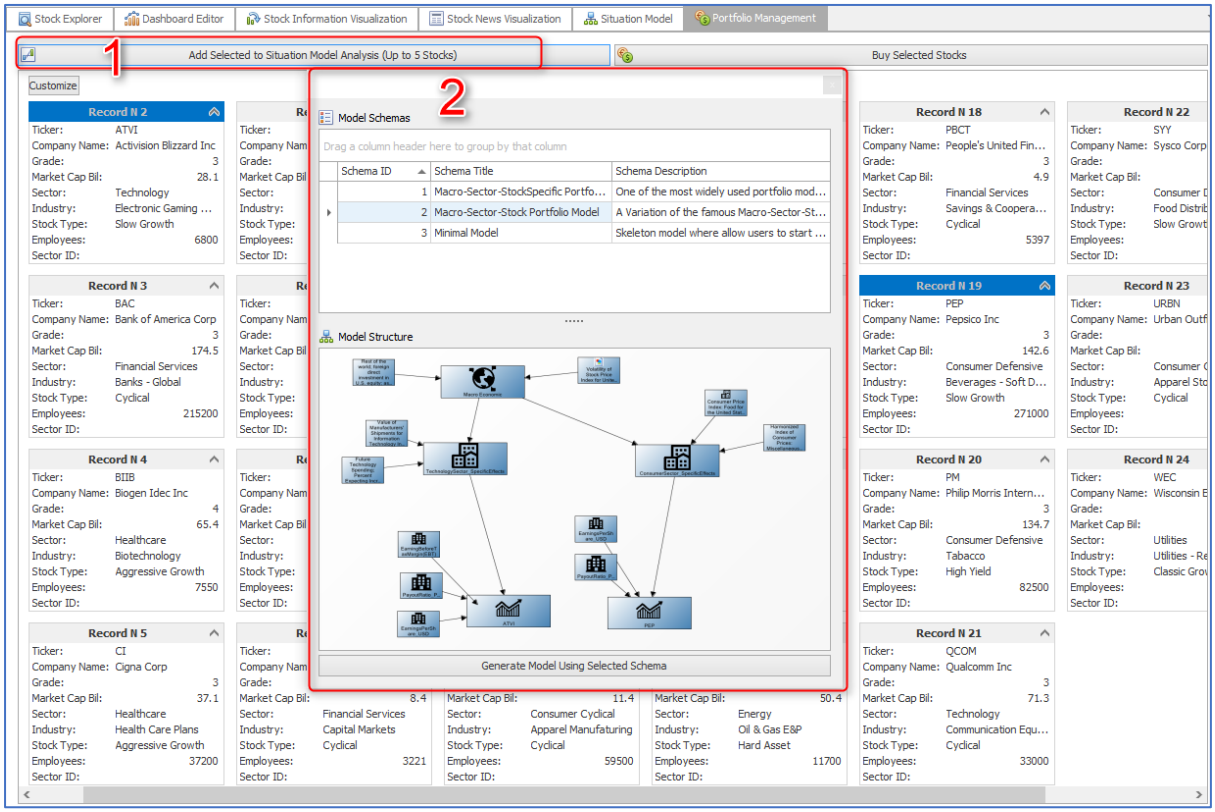


Figure 62. Enabling users to start the situation modeling with some established templates

As the actualization of the design "***Supporting situation modeling that can have both data-driven factors and argument-driven factors to build a situation model***", the design enables data analysts to build situation models from both quantitative and qualitative information. *Figure 63* shows the interface for situation modeling. To start building a situation model, users can select the data-driven factors from the "Quick Drop Blocks" on the left panel or can insert an argument or an association that they derived previously. This study encourages the users to use the data-driven factors whenever it is possible, whereas the qualitative-based argument and association are used as complement components for the situation model. This is because the data-driven factors use available data to derive the statistical inference about a key factor in the problem situation. However, human intuition often ignores the base value from the data. Therefore, the use of data-driven factors can enhance the accuracy of the situation model. Note that the number of factors in the network can cause a combinatorial explosion of the relationships and the complexity of the underlying equations. It can slow down the computation performance significantly. In the prototype, users are limited to at most 30 factors in a single network.

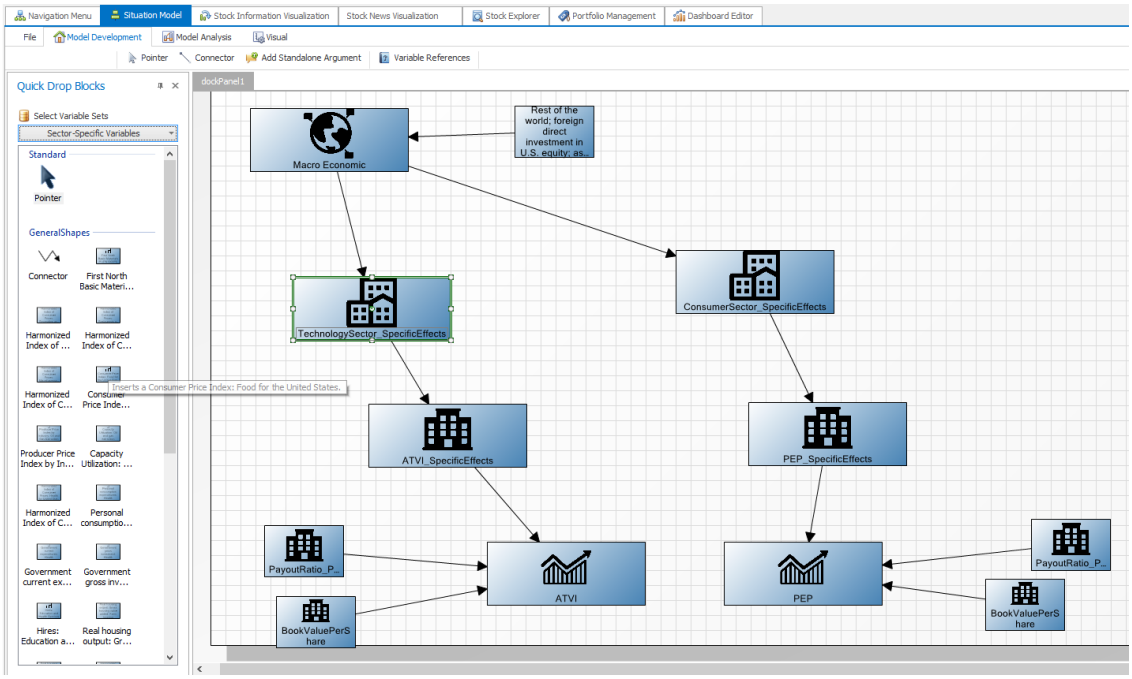


Figure 63. Interface for situation modeling

After building the structure of the situation model, users can turn on the situation model as a "dynamic situation model", as shown in *Figure 64*. During this transition, the prototype uses data to validate the factors and the links in the situation model wherever possible. Note that now the factors have become rectangles with horizontal bar charts within the rectangles. The bars in the chart represent the probabilities of different possible states of a factor. For data-driven factors, the possibility of the different states is computed based on the historical data of the factor. Taking "inflation rate" as a data-driven factor, the probability shows how the inflation rate would be based on historical data. For example, 43% of the chance at 3%, 50% of the chance at 5%, and so on. For the argument factor, the system derives the probability based on the confidence level and base value that the users entered.

---- This Space is Intentionally Left Blank -----

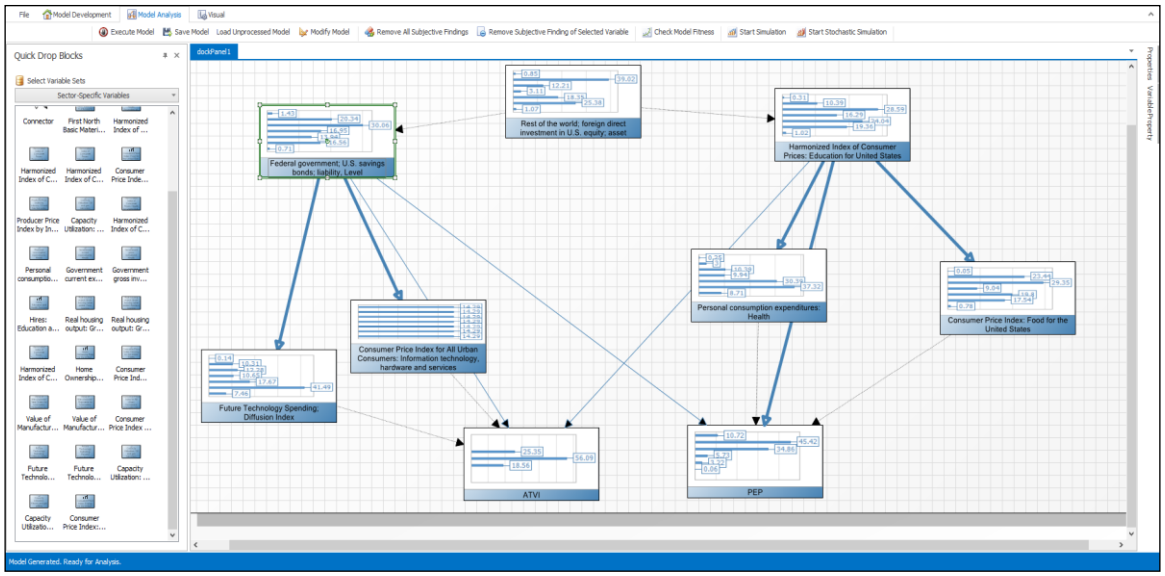


Figure 64. Dynamic situation model

Besides the factors, the system also establishes the influence value for the links between the factors as shown in *Figure 65*. Multiple regression technique is used to inform the analysts the level of confidence they can have that the association is true (i.e. statistical confidence level) and the strength of the association. Based on the statistical inference, the prototype will display all the positive links as blue, while negative links are in red. A positive link implies that the factor may have a positive correlation relationship with the other factor at the arrowed-end of the link. On the other hand, red-color link implies a negative correlation relationship. The thickness of the links represents the strength of the correlation.

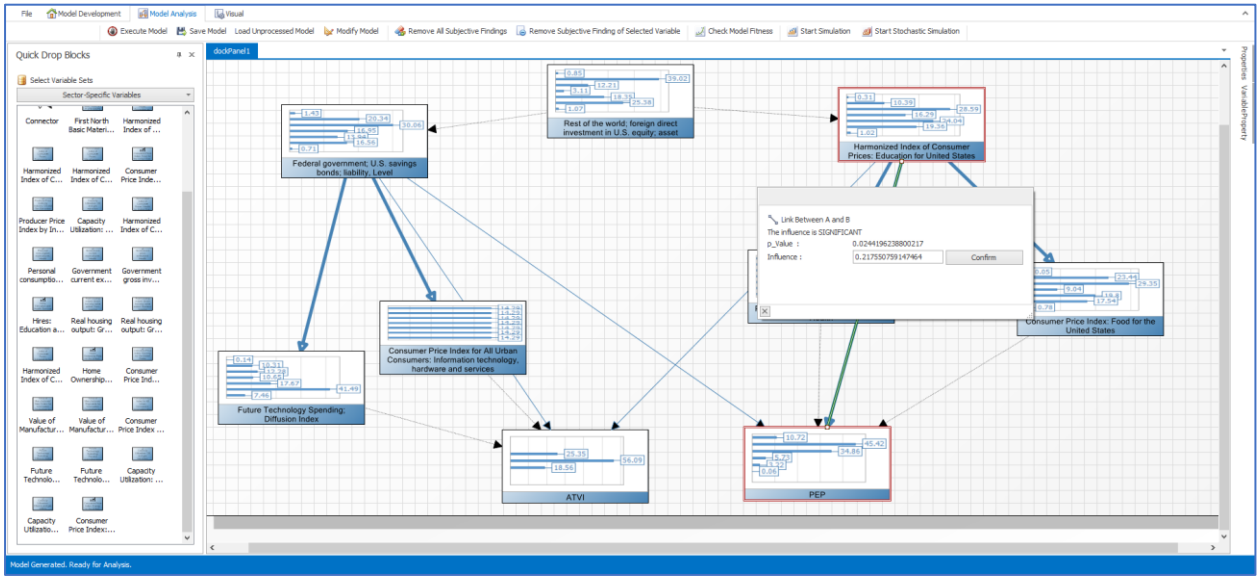


Figure 65. Relationships between the factors

Note that this “dynamic situation model” is generated based on the statistical inference from the data and the first-round of user inputs. It provides a basic view of how the factors in a situation model interrelated with each other. For instance, the models illustrate how macroeconomics factors influence the industry-specific factors, which in turn have impacts on the company stock price. All the influences are specified in terms of a probability distribution of the event or entity. The probability in the factor and the causal influence make the baseline value determined based on the historical data. However, in some cases, users might have a very clear, but different idea of how the situation model would be, rather than the one presented to them. Analyst can freely override the baseline value, based on their domain knowledge, judgement, or assumptions. They can fine-tune the situation model by adjusting the probability on the factors or the strength of the causal relationships.

The design “supporting a visual modeling technique for specifying the situation model that is interactive, dynamic, and computation-supportable” is actualized in the overall design of the situation modelling module, instead of by the specific functionalities. By visually creating the situation model, the users have actually created a complex mathematical model underneath the interface. The easy-to-use, intuitive interface has released the users from the complexity. The users do not have to understand how the probability calculation works and they do not have to specify the equation for testing the model against the data. Yet, now they have a complete casual probability network that can take advantage of computational power to do prediction, model validation, and more.

5.3.5.4 Intended Effects

It is alleged that with the designs in place, the users will be able to effectively build a situation model that can represent the big picture of the problem situation. Firstly, the design “to use established conceptual framework as the core structure of the situation model” allows the users to quickly start off a situation model with an effective structure. Without such a clue, users might spend significant time in searching the initial structure of model by repetitively building and scraping the structure until they find a satisfactory one. Therefore, the design reduces the time and effort required from the analysts by giving them a head start in the building process.

The second design enables users to flexibly use quantitative and qualitative information to build a situation model. Wherever the quantitative information is available, the users can directly use it as a building block in the situation model. This allows the users to build a situation model by taking the advantage of the objectivity of quantitative data. Whenever a qualitative argument is needed, the users can specify their own building block in the situation model. It allows the user to use the information without the needs to enter to the database beforehand. Users are able to incorporate the latest updates or news they have learnt into the situation model instantly for analysis, without the lead time for requesting changes in the database.

The third design is manifested in the overall situation modelling module. It enables the situation model to be represented as a computer-recognizable external representation. Compared to conventional data analytics systems, where users often have to create a situation model in their mind, the externally represented situation model can be the input and stimuli to the user's reasoning (Qu & Furnas, 2005). By facilitate explicit encoding of information, it reduces the cognitive strains on the users to mentally maintain the structure of the situation model. Moreover, visual-enabled situation modelling is close to natural language, which reflects the ways users talk and think about decisions. As the outcome, the design allows the users to focus on their analytical reasoning at the semantic-level of the model building, such as connecting the missing dots between the factors in the situation model and refining the model structure to more closely represent a problem situation in the physical world.

Considering the effects of these designs from the design principle "enabling assisted situation modelling", this study conjectures that a data analysis incorporating the design principle can help to enhance user performance in the connect & build activity during the information synthesis phase.

Proposition: The data analytics systems with capability for ***enabling assisted situation modeling*** will allow users to perform better in the *connect & build* activity.

5.3.6 Enabling Predictive Reasoning

5.3.6.1 Overview

- To support the modelling, representation, and storage of multiple hypothesized scenarios
- To support the prediction and simulation with the aids of computer-aided reasoning that can be flexibly steered by the users to reflect their intention, judgment, and knowledge.

This design principle "enabling predictive reasoning" stresses the importance for supporting the data analysts to predict and reason about the states of the problem situation. The objective of the design principle is to enhance the accuracy and efficiency of such predictive reasoning activity. The design principle contains an IS initiative, namely "enabling user-driven predictive reasoning that is complemented by advanced analytic techniques". With these design initiatives in place, this study conjectures that the human analysts can achieve more effective mental prediction and simulation which will help them to better understand the consequence of various scenarios.

5.3.6.2 IS Initiatives

The aim of the IS initiative "***enabling user-driven predictive reasoning complemented by advanced analytics techniques***" is to provide mechanisms that enable the data analysts to engage in an interactive process of predictive reasoning. As opposed to traditional prediction techniques that are rigid, the user-driven predictive reasoning in this study is different, in the sense that 1) its prediction target is the whole

situation model rather than just the specific dependent variables, 2) it allows users to engage in a continuous process, fine-tuning the prediction on the fly, and 3) it is a stochastic prediction that copes well with uncertainty and missing information.

Recall that prediction and simulation go hand in hand in analyzing a situation model. The prediction involves posing different "what if" questions. For instance, the users would like predict how the inflation rate will be. Then, the users try to mentally simulate how the inflation rate will influence other factors of the problem situation, based on the connection between the factors in the situation model. However, conducting the prediction and simulation entirely in the users' mind is highly inefficient. Researchers have shown that humans cannot reason effectively about scenarios that are unavailable to them (Chinchor & Pike, 2009; Heuer, 1999). It is also well proven that doing complex analysis primarily in one's head is more prone to various cognitive biases (Thomas et al., 1993). For these reasons, this IS initiative strives for the symbiosis between human users and advanced analytics techniques.

The IS initiative "supporting predictive reasoning which can be driven by human analysts and complemented by advanced analytics techniques" contains two important components. First is the "driven by human users" and the second is "*complemented by advanced analytics techniques*". Each of these components has different implications for the design.

Firstly, "driven by human users" implies that the predictive reasoning is an interactive process where users are actively steering or refining the process. To achieve this interactive process, the prediction and simulation need to be externalized. This study conjectures that as the complexity of the model increases, it becomes more rewarding to predict and simulate the model by constructing a physical representation. The physical representation allows users to engage in a deeper interaction between them and the problem situation, in order to understand how the problem situation would react to different stimuli or assumptions. Moreover, this physical working model enables the users to predict and simulate with greater precision for faster and longer times, compared to without such external representation (Kirsh, 2010). This external representation is made possible by the underlying mathematical models and statistical equations automatically generated when users created the situation model. Otherwise, huge efforts and technical knowledge will be required to develop a prediction model. More importantly, it enables the situation model to take advantage of machine-aided reasoning techniques to support the prediction and simulation of the users.

Moreover, "driven by human users" also implies that the users should take the lead in the analytical reasoning process. Predictive reasoning is a process that is inseparable from the human data analysts, especially for complex problem solving. It requires human judgement to make the best possible gauge of incomplete, inconsistent, and potentially deceptive information (Thomas & Cook, 2005). The uncertain and nonlinear relationships in a complex problem require the users to engage in iterative trial-

and-error-alike processes. Therefore, the supports for the predictive reasoning should also mimic the human way of constructing mental understanding: it is not a static one-off process but subject to continuous refinement as the users learnt in the current iteration of predictive reasoning. Specifically, the predictive engine should allow the analysts to incorporate newly learnt knowledge into the predictive model and to reflect the update instantly, without disrupting the reasoning flow.

The second component of the IS initiative, namely "*complemented by advanced analytics techniques"*, implies that the computer-aided techniques are used to directly enhance the users' predictive reasoning, instead of just relieving the load on the users' working memory and attention span. In a simple problem that involves very few factors, users may easily generate a few "what if" questions of a single factor and try to simulate the effects on other factors. However, in a complex analytics problem, this is almost impossible because 1) there are too many factors and each factor can be subject to many "what if" questions, 2) the factors are complexly interrelated, and 3) the users may not know which the useful "what if" questions for certain factors are if they are not well acquainted with the factor. For example, users can easily create a few "what if" questions based on *inflation rate,* such as what if the inflation is 2.5%, 4%, or 5%. This is because the users know the historical trend of the inflation rate. More importantly, the factor has less uncertainty or less fluctuation. Nevertheless, in a complex analytics problem, there are many factors which the users are not acquainted with or factors that are highly fluctuated. Therefore, it is important to use computer-aided predictive or forecasting techniques to suggest to users which are the best "what if" questions to ask.

Additionally, mentally simulating the effects of the "what if" question on a highly-connected situation model is nearly impossible, particularly when stimulating multiple "what if" questions at the same time. As such, the data analytics systems must carry out the simulation on behalf of the users. This study uses the Bayesian network as the reasoning engine for predicting the effects of the "what if" questions on the situation model as a whole. Comparison studies have shown the performance of Bayesian network to be superior to other unsupervised techniques such as Artificial Neural Network and Decision Tree which do not incorporate domain knowledge from the users (Lee & Chang, 2009). The power of the Bayesian network lies in its ability to incorporate domain knowledge, which makes it suitable for a situation model that is uncertain and subject to missing data. Many people misunderstood that causal probabilistic network only work well if the probabilities that the network based on are highly accurate. This has been found not to be true, and often that approximate probabilities, even subjective ones that are a guess, give very good results. Often the combination of several strands of imperfect knowledge lead to surprisingly strong predictions (Demirer et al., 2006).

Therefore, this study proposes that a semi-automatic predictive engine driven by human reasoning would be able to facilitate the predictive reasoning in complex problems. This study believes that this method can take advantage of the machine computation to process huge amount of data in order to

unveil potential predictions rigorously, while the data analysts can connect the missing dots or override the predictions, based on their domain knowledge and heuristic judgment, to iteratively refine the predictions. *Figure 66* shows the conceptual illustration of the predictive reasoning.

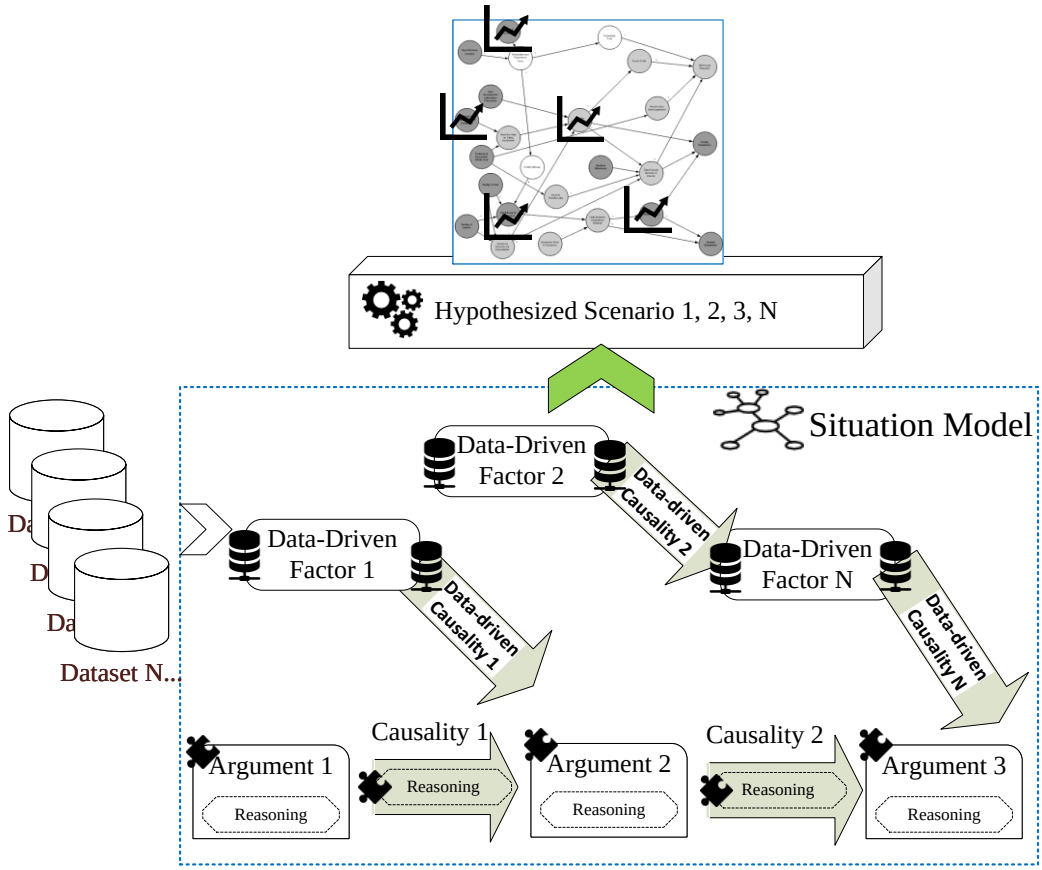


Figure 66. Supporting predictive reasoning

5.3.6.3 Mechanism

Two aspects of predictive reasoning need to be supported to enhance the performance of analysts during the predictive reasoning phase. Firstly, given that the human's shortcoming is in reliably carrying out the prediction of a complex system, machine-driven predictive techniques can be used to increase the accuracy and to reduce the chance of biases. Secondly, the mental simulation in a complex problem situation requires the analysts to quickly try out different assumptions on the fly to obtain feedback; the feedback is potentially used to guide or configure the next mental simulation.

Recall that design principle "assisted situation modeling" allows users to generate a "dynamic situation model". The dynamic situation model can be used by users for predictive reasoning. Users can first predict or pose a "what if" question using the situation model. In *Figure 67*, the sliders on the right panel can be used to incorporate users' prediction into a particular factor in the situation model. Users

can adjust the slider to represent how the probability of different states would be in the future. The adjustment can represent the user's assumptions or speculation about a particular factor. It is important to augment a representation with uncertainty in order to allow potential interpretations of the data to be considered (Zuk & Carpendale, 2007).

As previously discussed, if the users are familiar with the factor, they can confidently adjust the probability manually. Besides using the sliders, users can also enter the range of the factor's value as their prediction. For example, users can enter 2.5% and 3.5% to represent the range of inflation rate that the users think the actual inflation rate will be.

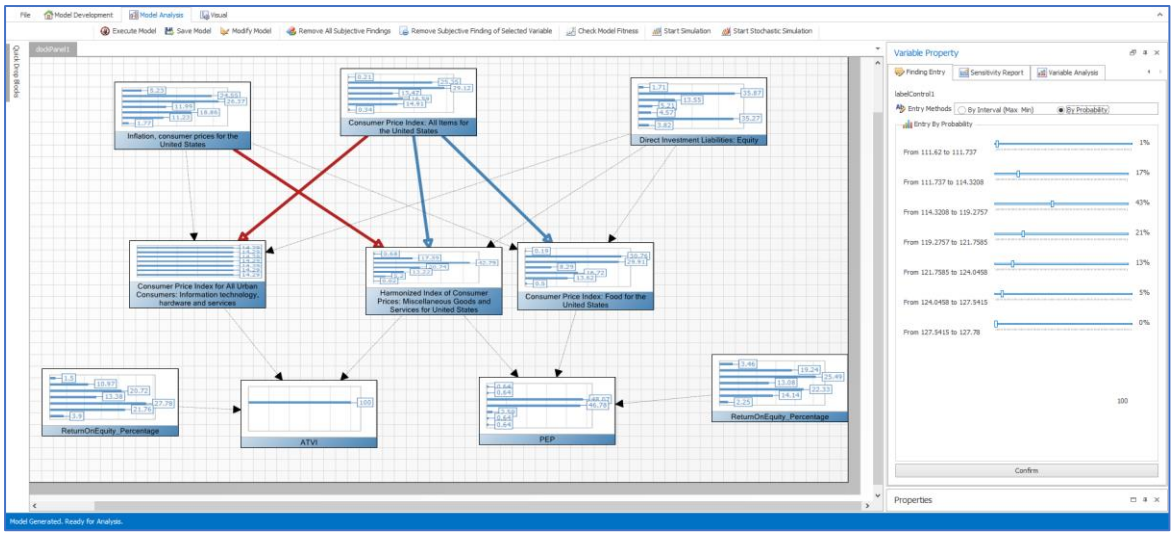


Figure 67. Enabling prediction or posing what-if question

If the users are not familiar or not confident to predict the states of a particular factor, they can use one of the built-in forecasting algorithms to facilitate their prediction. *Figure 68* shows the interface from which the users can choose and preview the prediction resulted from the different forecasting algorithms. Note that this aid is available for only the data-driven factors. Once the predictive algorithm is chosen, the result will be used to specify the probability of the states of the factor. Thus, the prediction of a factor is entered.

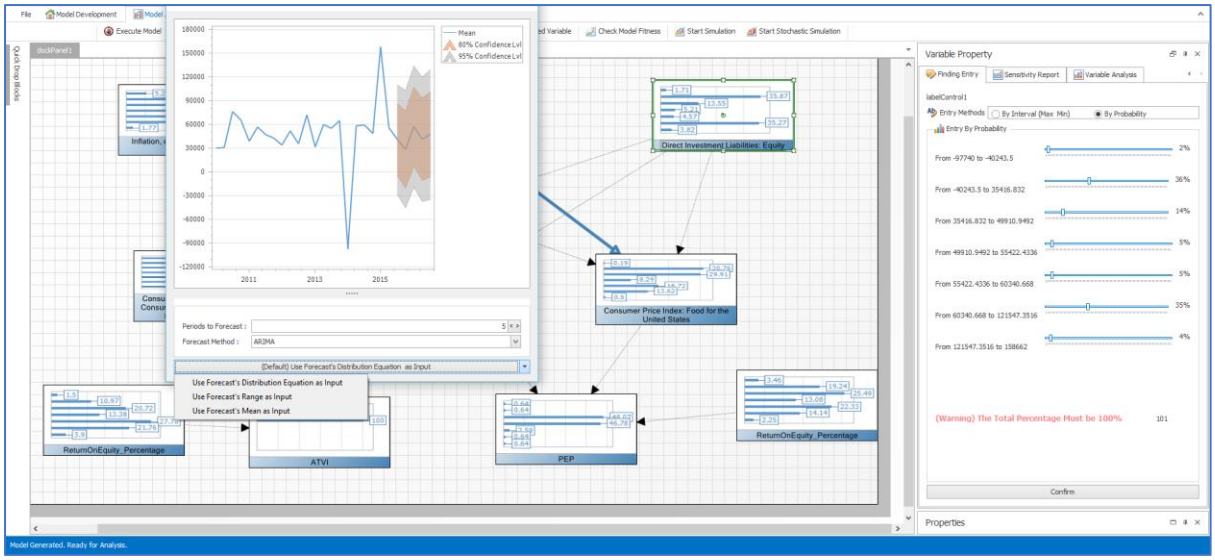


Figure 68. Prediction facilitated by forecasting algorithms

Every time that users enter the prediction of a factor, the dynamic situation model will instantly simulate the effects of the prediction on the other factors in the model. Precisely, the situation model automatically propagates the effects of the prediction to the entire model based on the links between the factors. This provides instant feedback to the users about their prediction. More importantly, this feature is useful for the users to distinguish among the effects of multiple predictions within the same model.

5.3.6.4 Intended Effects

The designs from the design principle "predictive reasoning" turn the situation model into a working model which users can interact with to predict the consequences of different assumptions and speculations. Compared to the static external representation, the dynamic model allows the users to shift the focus onto the reasoning and simulating of the situation model, thus allowing them to be better acquainted with the situation model. The dynamic representation changes the role of the users from a "sense and respond" focus to a forward-looking "predict and anticipated" focus. It allows the users to predict how the system would respond to different stimuli or conditions, helping the data analysts to understand the dynamics of the problem situation in order to prioritize their attention.

Moreover, the design takes advantage of the computer-aided prediction algorithm to aid the prediction of the future states of the situation. It encourages rigorous and logical processing which is able to enhance the validity of the reasoning outcomes and to reduce cognitive pitfalls. Moreover, the predicted situation model is presented with the uncertainties inherited from each of the factors within the situation, thus providing cues about a) uncertainty to promote transparency of assumptions and b) weak points in the predicted situation model (Zuk & Carpendale, 2007).

Overall, this study conjectures that those data analytics systems that have the capability to enable predictive reasoning can enhance user performance in the predict & simulate activity during the knowledge actualization phase.

Proposition: The data analytics systems with a capability for ***enabling predictive reasoning*** will allow users to perform better in the *predict & simulate* activity.

5.3.7 Enabling Stochastic Optimization

5.3.7.1 Overview

- **Design Requirement:** To support users in optimizing the resource allocation that can meet the conflicting objectives within their constraints, while compensating for the risks.

- **Design Requirement:** To Support users in accurately and rigorously assessing the risks associated with the courses of action.

The design principle "enabling simulative optimization" emphasizes the enabling of the data analysts to formulate an optimal resource allocation plan. The resource allocation plan is optimized to meet the conflicting objectives, while being compensated for the uncertainty in a hypothesized scenario. The design principle consists of two IS initiatives, namely 1) supporting user-driven optimization and 2) enabling user-driven risk assessment. This study conjectures that the design principle enables the data analysts to rigorously evaluating the course of actions in the light of their objectives, constraints, and uncertainty. As the result, the analysts can better understand the risk involved in their course of action and are better informed whether they should actually execute the planned courses of action.

5.3.7.2 IS Initiative

The IS initiative "***supporting user-driven optimization***" is important for the complex problem situation. Solving a complex problem requires the users to decide which action is to be taken. In turn, to take action often requires resource allocation in order for the action to be effective and efficient. The resource allocation is very important as it significantly determines whether an action plan will succeed or fail. The resource allocation determines how well the action plan can meet the objectives within the limited resource, without not violating any constraint. Even a brilliant action plan would fail if the resource allocation for the plan was poor. Therefore, there is great need to optimize the resource allocation after the users have identified their potential action plans.

The basic optimization involves making the most effective use of the resource allocation, against conflicting objectives, limited resources, and constraints. Even with very simple optimization, the users need a computer or at least paper and pen to complete the calculation. Increase in the numbers of action plans, objectives, resource types, and constraints, soon require dedicated computation algorithms to do the optimization. For solving a complex problem, an additional factor to consider during the optimization is the uncertainty inherent in the situation model. Recall that all the factors in a situation

are uncertain, as they are described by the probability of its plausible states. For instance, after the predictive reasoning process, users have identified two profitable stock options, namely stock A and stock B. The users speculated that the prices of these stocks will increase by the chances shown in *Table 14*.

Table 14. A simple example of uncertainty in a situation model's factors

Stock A		Stock B	
Increase Rate	Chance	Increase Rate	Chance
5.0 to 7.5 percent	25%	3.1 to 6.0 percent	50%
7.6 to 10.0 percent	45%	6.1 to 9.0 percent	15%
10.1 to 12.5 percent	15%	9.1 to 12.0 percent	15%
		12.1 to 15.0 percent	20%

As a result, the optimization also needs to factor in the uncertainty in the factors. The optimization needs to maximize the users' objectives while minimizing the risk due to the uncertainty. In the stock analysis case, optimization needs to maximize the return of investment (ROI) while minimizing the risk of loss due to the uncertainty. A stochastic optimization is needed for optimizing against uncertainties. For the optimization process, users should be given the flexibility of experimenting with the variables that they control. For example, the system should support users in experimenting with different risk profiles and preference on choosing stocks. The repetitive experimentation with different variables allows the users to observe and learn how these different controllable variables could affect their objectives.

Once completed, the optimization mechanism should present an optimized resource allocation plan to the users. When using a user-driven design, the users should have the full authority whether to adopt the plan, adjusting the plan based on domain knowledge, or ignore the plan. Note that the quality of the optimized resource allocation plan is only as accurate as the situation model specified by the users. This points out the importance for the users to assess the "fitness" of their situation. It is not just for the users to gain accurate assessment about the problem situation; but the situation model also determines the quality of the optimized resource allocation plan. From the other perspective, the optimization is unique for every user because the optimization takes their uniquely built situation model as the inputs. In other words, the optimized resource allocation plan is uniquely tailored to how the particular user perceives the problem situation.

Design: Supporting users in optimizing the resource allocation
that can meet the conflicting objectives within their constraints,
while compensating for the risks

The IS initiative "***supporting user-driven risk assessment***" is important because the users' decision whether to follow the optimized plan depends on how confident they are about the plan. They will need to know what chance there is that things do not happen as expected. This calls for the need to enable the users to assess risk associated with the optimized plan. In predictive reasoning, the users can pose a "what-if" question at a time and observe the impact. This approach is practical only for the users to test a small number of speculations or assumptions. The risk assessment involves exhaustively testing all possible "what if" questions. The virtually unlimited possibility requires computer-aided technique for the processing. The computer support should automatically generate all the possibilities based on the probability distribution associated with the factors in the situation model. It should then present the associated risk in a way to facilitate users' understanding. This IS initiative is aligned with the visual analytics community's call for the design of machine processing and visualization to complement the human analyst's ability to understanding uncertainties (Keim et al., 2010). *Figure 69* shows the conceptual illustration of the enabling optimization and risk assessment.

Design: Supporting users in accurately and rigorously assessing the risks associated with the courses of action.

---- This Space is Intentionally Left Blank -----

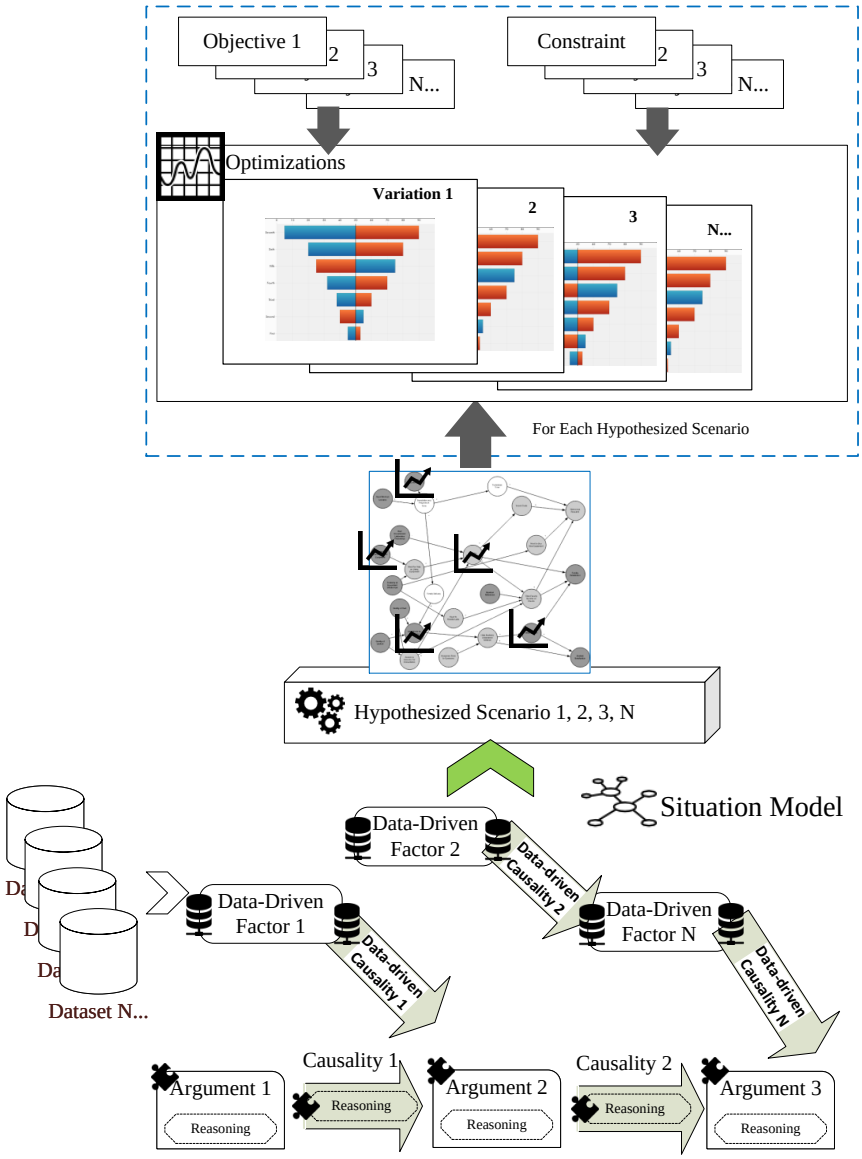


Figure 69. Enabling optimization and risk assessment

5.3.7.3 Mechanism

Once the users are satisfied with the situation model, they can invoke the optimization and risk assessment mechanism. Both the designs are actualized on a single system feature because the optimization and risk assessment in practice are closely interrelated. *Figure 70* shows the interface for the optimization and risk assessment mechanism. Once invoked, the optimization will automatically start. The resulting key information about the optimization is shown in Box 1. *Figure 71* shows the enlarged view of the information, which contains the optimized resource allocation plan. In this particular example, the resource allocation plan is the optimal number of stocks to invest: namely, the stock portfolio. In order to understand the risk associated with this plan, the users can refer to Box 2. The graph shows the probability distribution of the average returns of the stock portfolio. Box 3 shows

the probability distribution of the individual stock options which the total return is based on. If interested, users can expand the panel to have a more detailed view of these probability distribution graphs.

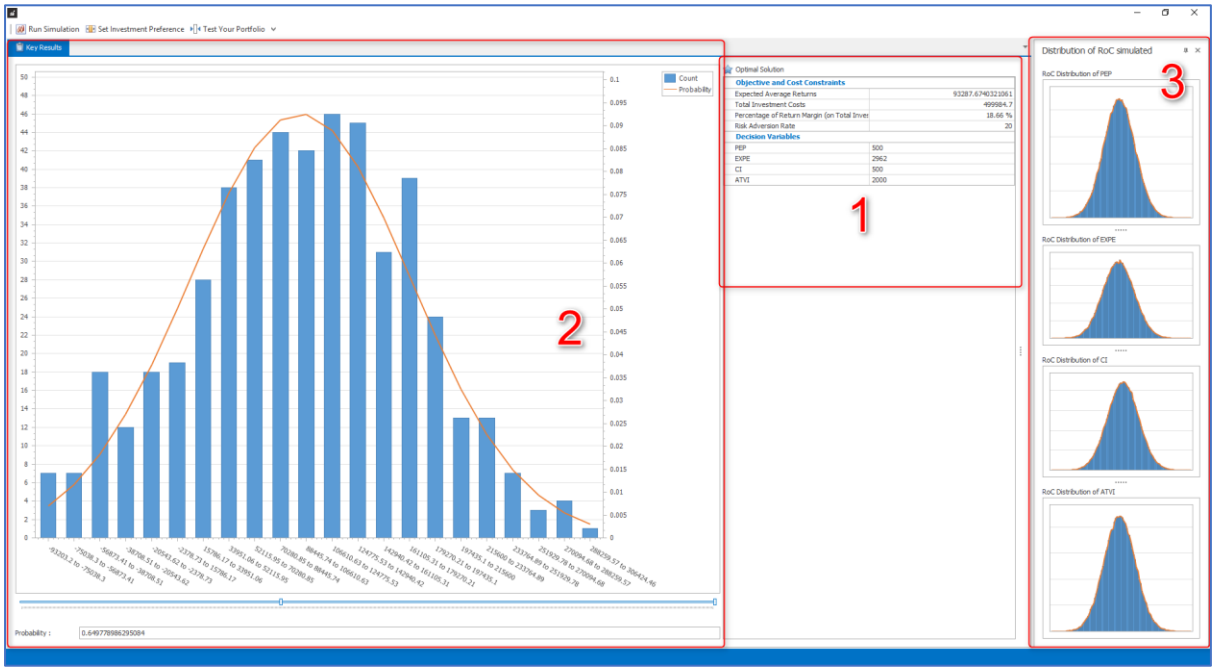


Figure 70. Interface for optimization and risk assessment

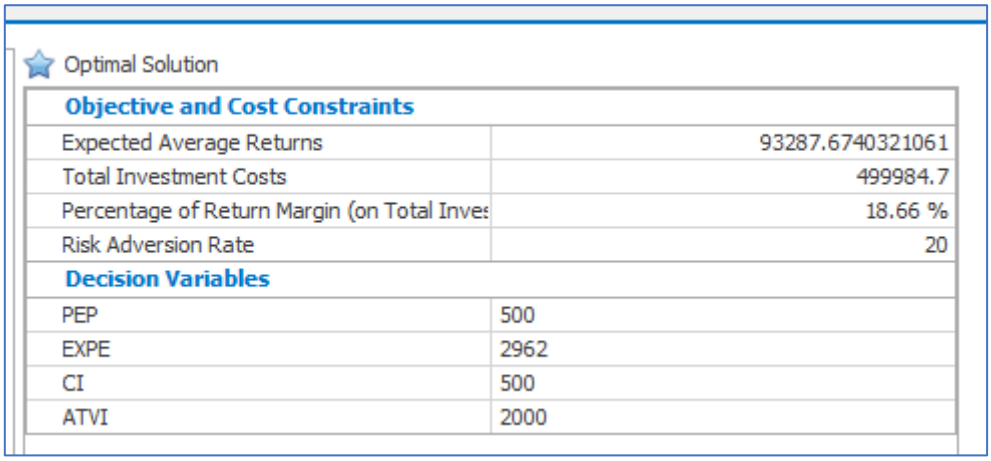


Figure 71. Enlarged view of the key information in box 1

In Box 2, in order to facilitate the understanding of the risk, users can use the slider or the input field at the bottom to interact with the risk assessment. *Figure 72* shows the enlarged view of the region that contains the slider and input field. The two interaction mechanisms interact in two ways. For example, the users can use to slider to indicate the range of total returns (in US dollars). The users in this particular case want to know what the probability of the portfolio that will result $86,503 or more. In other words, the users want to how likely is it that the portfolio will result in a 17% margin of return.

While the users move the slider, the probability will be instantly updated. Likewise, the users can enter a probability as a confidence level, and see which range of returns will fall into that confidence interval. For example, users might want to know what the range of return is that will fall into the confidence interval of 90%, 95%, and 99%. The interaction between the users and the risk assessment information will allow the users to better understanding the risk associated with the portfolio, and thus will allow them to make confident decisions whether to invest according to the suggest portfolio.

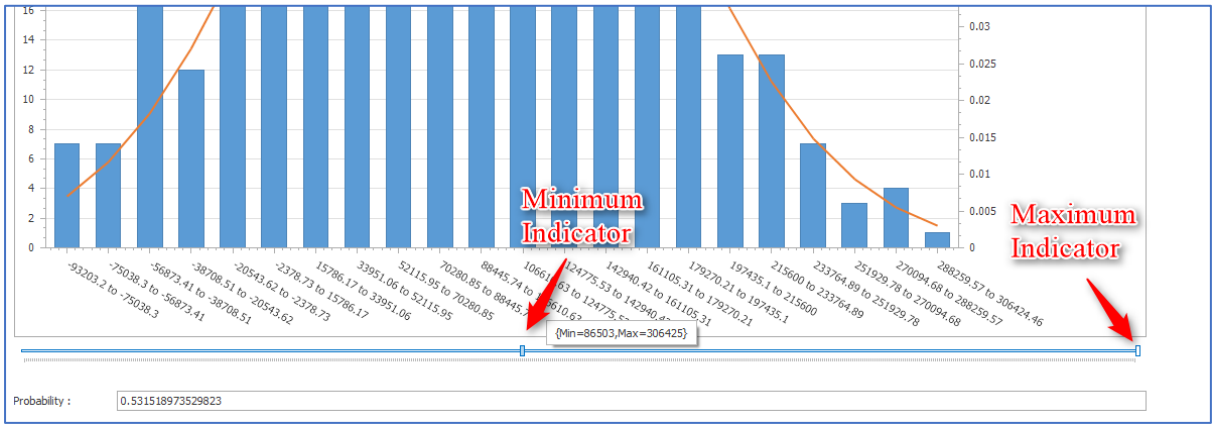


Figure 72. Slider and input field for interaction

As users would wish to experiment with different controllable variables, the purpose is to allow them to observe and learn how their action or preference could affect the objectives. *Figure 73* shows the pop-up dialog where user can experiment with different variables. The risk adverse ratio refers to the user's risk profile about what percentage of the capital they would risk to lose. The users can also specify their preference on the number of stocks to invest.

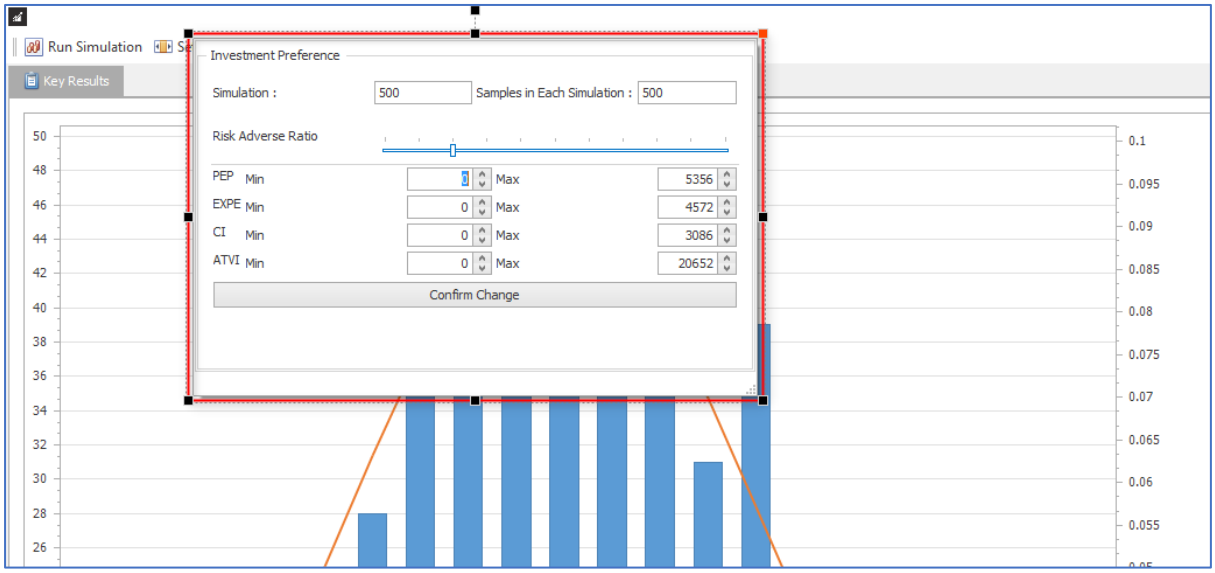


Figure 73. Pop-up dialog for tweaking the optimization and risk assessment

Very often, users will adjust the suggested optimized resource plan according to their knowledge and experience from the domain. In this example, once the users customize their own stock portfolio, they can use the dialog to specify the portfolio and assess the risks associated it.

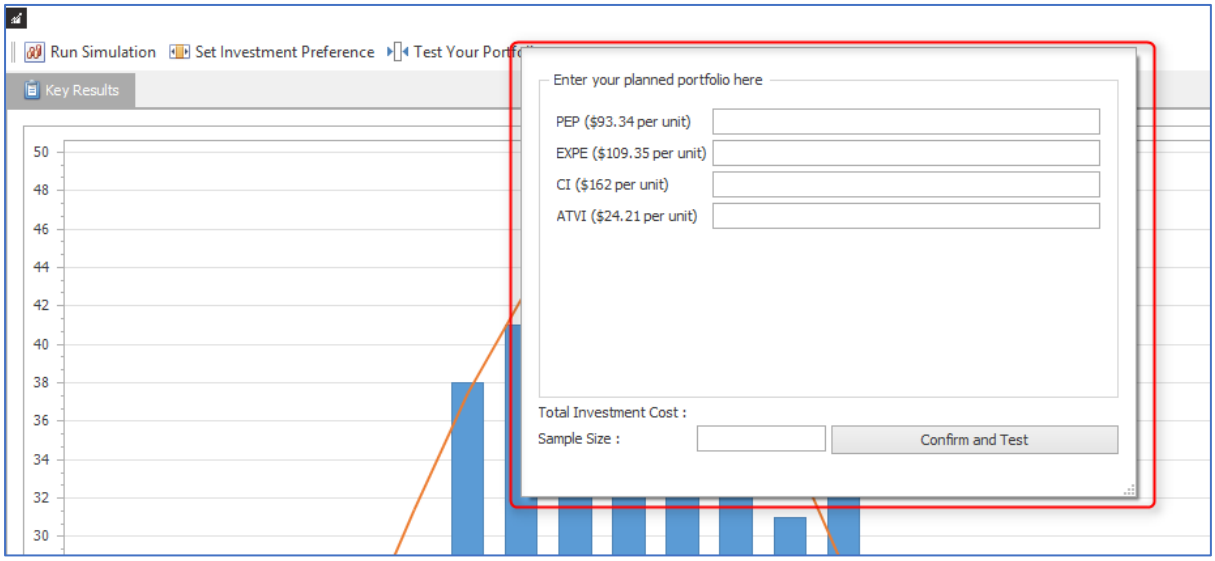


Figure 74. Enabling users to assess the risk of their own resource allocation plan

The features have been presented are enabled by the stochastic simulation under the hood. The simulation takes the advantage of Monte Carlo simulation and linear programming to optimize the funds allocation. The stochastic simulation module receives the outputs from the probability network in the form of the probability distribution of the selected stocks. The probability distribution is used to generate the samples that are subsequently used to construct different possible scenarios of the stock prices. Then, a recursive linear programming is applied to calculate the fund allocation that can maximize the return of investment while minimize the variability in the scenarios.

5.3.7.4 Intended Effects

The designs from the "enabling stochastic optimization" design principle facilitate the optimize & assess risk activity in two ways.

Firstly, the design allows the users to be aware of the resource allocation plan that can best meet their objectives within the constraints that they have. The users can experiment with variables that they have control over in the real world and to learn the potential effects of different combinations of these variables. The kind of trial-and-error learning is hardly able to be conducted in the users' mind, due to the complexity and the interconnected objectives and constraints. Likewise, such learning in the real-world will be much too costly in terms of time and resource. Therefore, the design allows the users to learn and plan for their courses of action in a very effective way.

Secondly, the design allows the users to systematically assess the risk associated with their courses of action. Most importantly, the risk being computed based on the situation is tailored against how the users perceive that the problem situation works. By using a statistical approach to analyze and representing the risk to the users, the design can help to alleviate the issues with the biases and overinflated confidence. Therefore, this study conjectures that data analytics systems that are capable of enabling stochastic optimization can enhance the user performance in the optimize & assess risk activity.

Proposition: The data analytics systems with capability for ***enabling stochastic optimization*** will allow users to perform better in the *optimize & assess risk* activity.

5.4 Conceptual Design Framework

As a whole, the design, which combines the strengths of machines with the strengths of human analysts for enhancing the human analytical reasoning capability, allows the data analysts to effectively carry out the problem-solving activities in data analytics. This human-machine symbiosis makes the system useful for solving complex analytics problems. Following the "machine-augmented cognition" design philosophy, the design emphasizes the inclusion of active human interaction into the data analysis process in order to combine flexibility, creativity, and domain knowledge with the computational power of today's computer.

In terms of user interactions, the overall design supports both inductive and deductive approaches. For an inductive approach, data analysts can follow the flow discussed in the design principles, starting off with data exploration and ending with knowledge actualization. For deductive approach, data analysts can first build their situation model, then collect evidence from the data exploration to support their speculations about the problem situation. Note that the discussion of the design principles uses the inductive workflow because it is relatively easier to be understood based on its incremental nature. The design also supports the deductive approach.

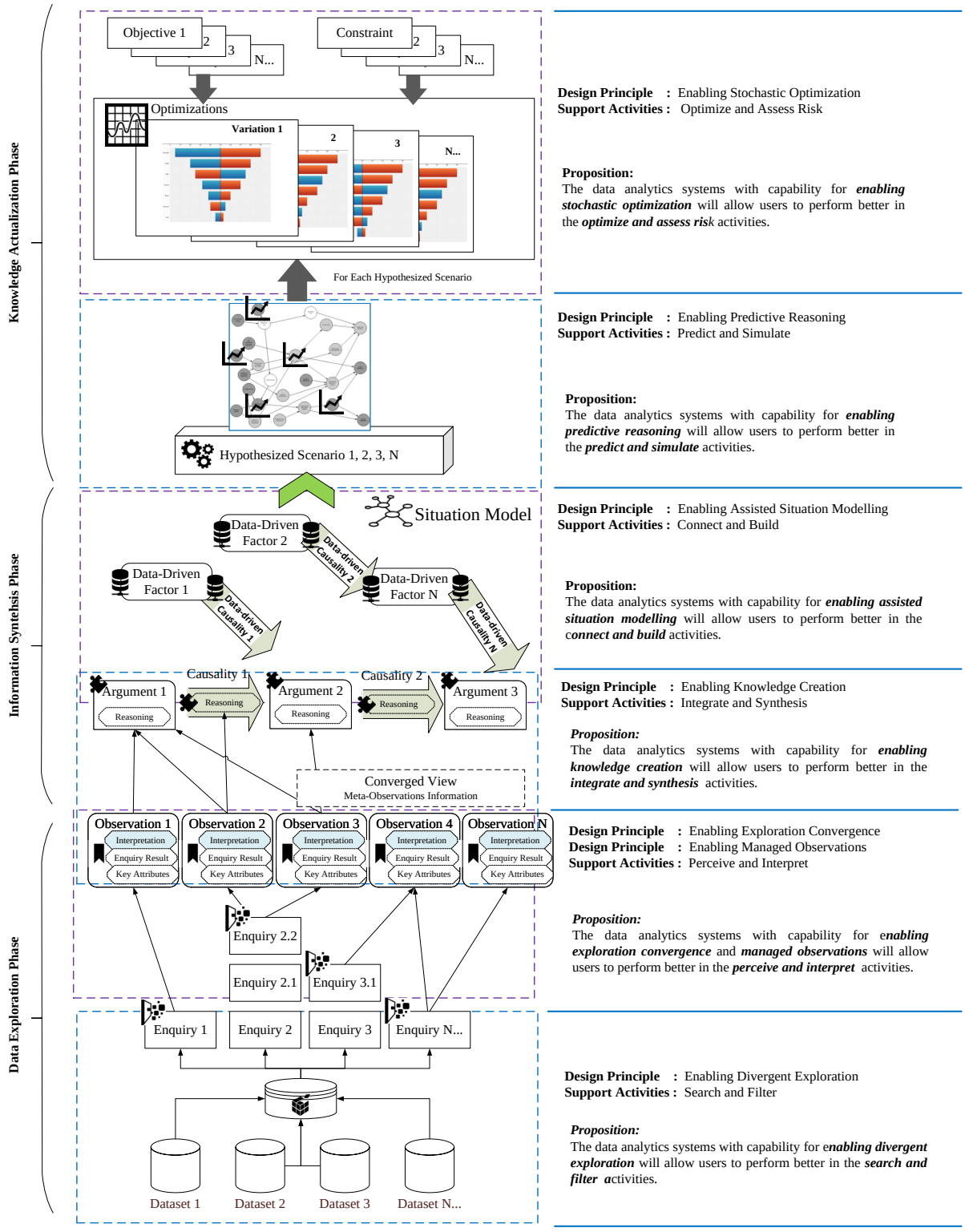
In practice, data analysts do not follow through a linear workflow of either inductive or deductive approaches. The data analysts often switch between the two approaches opportunistically. Moreover, they often engage in an iterative process between the phases of data analytics. The design is developed to cope with this nonlinear workflow. For example, while building the situation model, data analysts can easily switch back and forth between situation modeling and data exploration to find more evidence in order to continuously improve their situation model. Or better still, the design allows the data analysts to work on the data exploration, information synthesis, and knowledge actualization simultaneously. The analysts can access all the main interfaces at the same time, as long as they have sufficient space on their monitors. This is made possible by ensuring each of the main interfaces is programmed to be

run as a different process and to take advantage of multithreading architecture. In short, the design supports the fluidity of the sensemaking and reasoning of the human analysts.

One key capability enabled by the proposed data analytics system is that every data analysis is unique. Each data analytics is tailored to the situation model that the analysts created. The situation model works as a knowledge repository that reflects the experience gained from the analyses over time and the domain knowledge gained from the experts. This capability results in a retainable competitive advantage because an institute's power to the knowledge resides in the unique situation model that they have created, instead of in the software system that others can also easily acquire. The situation model become an intellectual asset in their data analytics process.

The overall design also enforces a concept this study called *persistence data analytics*, which allows the reasoning artifacts and other intermediate outcomes of the analysis to be stored and accessible. It offers two benefits: continuity of data analytics and facilitated communication and collaboration. This design is useful for complex analytical tasks because the data analysts often carry out the data analysis over a long period, ranging from days to months. The *persistence data analytics* design allows the data analysts to easily continue from where they left off in the previous session. Their stored reasoning artefacts such as observations, arguments, and the situation model work like a dynamic snapshot that can be restored to resume the analysis. In terms of facilitating communication and collaboration, the explicit representation of the reasoning artefacts surfaces the reasoning process, assumptions, and evidence that lead to the conclusion. It allows the data analysts to use the system as the presentation tool to explain their results to the decision makers. Such a capability bypasses the need for lengthy reports and presentations. Moreover, the multiple data analysts can collaborate to work on the same data analytics task. For instance, they can build the situation model together, each looking for the evidences from their area of specialization.

The design framework is a product of the integration of the design principles. *Figure 75* shows the overview of the conceptual design framework, with the right side of the figure showing the design principles and the corresponding problem-solving activities they support. The overall goal of the design principles aims to explicitly support the problem-solving activities in all the phases of data analytics, namely data exploration, information synthesis, and knowledge actualization. This study conjectures that by supporting the data analysts to effectively perform the problem-solving activities, will help them to achieve the insight components. The higher quality and the more complete insight components a data analyst possesses, the more likely it is that the data analyst will be able to gain actionable insight.



Design Principle : Enabling Stochastic Optimization
Support Activities : Optimize and Assess Risk

Proposition:
The data analytics systems with capability for ***enabling stochastic optimization*** will allow users to perform better in the ***optimize and assess risk*** activities.

Design Principle : Enabling Predictive Reasoning
Support Activities : Predict and Simulate

Proposition:
The data analytics systems with capability for ***enabling predictive reasoning*** will allow users to perform better in the ***predict and simulate*** activities.

Design Principle : Enabling Assisted Situation Modelling
Support Activities : Connect and Build

Proposition:
The data analytics systems with capability for ***enabling assisted situation modelling*** will allow users to perform better in the c***onnect and build*** activities.

Design Principle : Enabling Knowledge Creation
Support Activities : Integrate and Synthesis

Proposition:
The data analytics systems with capability for ***enabling knowledge creation*** will allow users to perform better in the ***integrate and synthesis*** activities.

Design Principle : Enabling Exploration Convergence
Design Principle : Enabling Managed Observations
Support Activities : Perceive and Interpret

Proposition:
The data analytics systems with capability for e***nabling exploration convergence*** and ***managed observations*** will allow users to perform better in the ***perceive and interpret*** activities.

Design Principle : Enabling Divergent Exploration
Support Activities : Search and Filter

Proposition:
The data analytics systems with capability for e***nabling divergent exploration*** will allow users to perform better in the ***search and filter a***ctivities.

Figure 75. Conceptual design framework

As a closing for this chapter, *Table 15* summarizes the key information in the design principles.

Table 15. Summary of the design principles

Design Principle: Enabling Divergent Exploration	
Design Requirement	▪ To support the users to effectively explore a large number of data elements.
Design Initiative	▪ Enabling divergence at data layer ▪ Enabling divergence at enquiry layer
Designs	▪ Supporting multiple datasets to be integrated to create a centralized data source for enquiry generation ▪ Supporting the generation of multimodal enquiries for perceiving multi-facets of the data
Intended Effect	▪ Support the search & filter activity
Design Principle: Enabling Managed Observation	
Design Requirement	▪ To support the users to capture, manage, and retrieve their observations, including the underlying interpretations
Design Initiative	▪ Enabling observations to be captured ▪ Enabling interpretation to be captured
Designs	▪ Supporting the observations and its enquiry context to be systematically captured ▪ Supporting the interpretation to be captured in structural form and be analysis-ready
Intended Effect	▪ Support the perceive & interpret activity
Design Principle: Enabling Exploration Convergence	
Design Requirement	▪ To support users to create joint summary from their observations
Design Initiative	▪ Enabling the meta-observation information
Design	▪ Supporting the visualization and analysis of the meta-observation information
Intended Effect	▪ Support the perceive & interpret activity

Design Principle: Enabling Exploration Convergence	
Design Requirement	▪ To support the users to create, manage, and retrieve high-level knowledge based on low-level analytic insights and reasoning
Design Initiative	▪ Enabling observation integration ▪ Enabling synthesis between observation and user reasoning
Designs	▪ Supporting the creation of high-level knowledge by integrating observations ▪ Supporting the user reasoning used to create the high-level knowledge to be structurally captured, stored, and retrieved
Intended Effect	▪ Support the integrate & synthesize activity
Design Principle: Enabling Assisted Situation Modelling	
Design Requirement	▪ To support the users identifying a preliminary structure of the situation model ▪ To support the users in using quantitative and qualitative information to build the situation model ▪ To support the users in constructing interactive, dynamic, and computation-friendly situation models
Design Initiative	▪ Enabling the selections of the core structure ▪ Enabling both quantitative and qualitative approaches to situation modelling ▪ Enabling dynamic situation model
Designs	▪ Supporting the use of established conceptual framework as the starting template of a situation model ▪ Supporting the use of both data-driven factors and argument-driven factors to build a situation model ▪ Supporting a visual modeling technique for specifying the situation model that is interactive, dynamic, and computation-supportable
Intended Effect	▪ Support the connect & build activity
Design Principle: Enabling Predictive Reasoning	
Design Requirement	▪ To support the modelling, representation, and storage of multiple hypothesized scenarios ▪ To support prediction and simulation with the aids of computer-aided reasoning that can be flexible steered by the users to reflect their intention, judgment, and knowledge
Design Initiative	▪ Enabling user-driven predictive reasoning that is complemented by advanced analytics techniques
Design	▪ Supporting a semi-automatic prediction and simulation engine driven by human reasoning
Intended Effect	▪ Support the predict & simulate activity

Design Principle: Enabling Stochastic Optimization	
Design Requirement	▪ To support users in optimizing the resource allocation that can meet the conflicting objectives within their constraints, while compensating for the risks. ▪ To support users in accurately and rigorously assessing the risks associated with the courses of action.
Design Initiative	▪ Enabling user-driven optimization ▪ Enabling user-driven risk assessment
Designs	▪ Supporting users in optimizing the resource allocation that can meet the conflicting objectives within their constraints, while compensating for the risks ▪ Supporting users in accurately and rigorously assessing the risks associated with the courses of action.
Intended Effect	▪ Support the optimize & assess risk activity

---- This Space is Intentionally Left Blank -----

Chapter 6
Designing the Evaluation

6.1 Overview of Evaluation

This chapter describes the research activity "designing the evaluation" introduced in *subsection 3.4*. The purpose of this research activity was to design a rigorous evaluation process for collecting the data, which in turn was used to test the propositions developed in the previous chapter. This activity includes 1) operationalizing the constructs of interest into measurable variables and 2) designing the data collection process, which both make up the main contents of this chapter. This chapter commences with Section 6.2 describes the requirements for the evaluation and this study's rationales in addressing the requirements. Based on these requirements, section 6.3 then shows the operationalization of the constructs, and section 6.4 describes the details of the data collection process, which includes participant recruitment, the procedure for the user study, and the design of the main task in the user study.

6.2 Requirements for the Evaluation

6.2.1 Needs for measuring the actual performance

The evaluation of data analytics systems has proved to be a challenging task. One of the challenges is attributed to its high-level goal to solve complex problems in practice. As a complex problem-solving process, there is no clear-cut criterion by with to judge whether the analysis outcome is good or bad, or whether the users have performed better or worse. To avoid this complication, some data analytics studies have chosen to measure user performance by counting the new data discovery the users have made (Saraiya, North, Lam, & Duca, 2006; Smuc et al., 2008). A data discovery can be a pattern, an association, or significant clusters. However, this study argues that such low-level measurement does not reflect the real goal of which the users using the system. In short, more information discovered does not necessary lead to better problem solving.

Merely measuring the final consequence of the decision can be misleading and does not fully reflect the analytical performance. A typical example of this measurement is the monetary value resulting from the decision. Such a post-factual performance measure is susceptible to other confounding factors that are beyond the system usage, such as how effective the plan being executed is, and the unpredictable random events. Therefore, such measurement may capture many other "noises" rather than the system effects. Moreover, analytical performance is not just about the quantifiable end outcomes. The total performance should also take into consideration the qualitative and intermediate products, such as the ability of the system in enabling users to engage in deeper reasoning, to consider more solution alternatives, and to learn to solve similar problems more effectively in the future.

In order to provide a more complete measurement of the analytical performance, this study measures the performance at three different stages of data analytics, as shown in *Figure 76*. The left side of the figure shows the analytical performance that reflects direct outcomes of the interaction between the users and the system. Toward the right end of the figure, the performance is closer to the high-level goal of data analytics systems. This performance is commonly how the end-users or organizations gauge the pragmatic value of the systems. However, more "noise" or confounding factors are factored in the performance. This three-stage performance evaluation allows this study to understand how the design effects traverse from the immediate implications of the system to a higher pragmatic implication that adds real value to the domain.

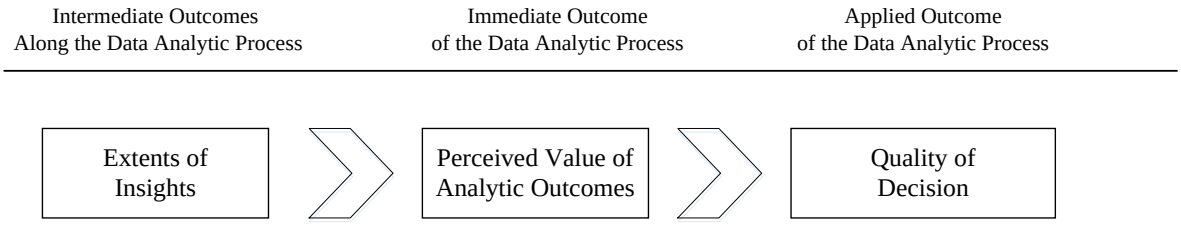


Figure 76. Performance at three different stages

Table 16 presents the three different stages and their corresponding constructs. *Extents of insight* measures a set of knowledge states the users gained about the analytic problem. It measures the user performance in the different problem-solving activities. This construct alleviates the issues of measuring the number of new data discoveries. It measures the users' discoveries by their significance relevant to the analytics problem. Supported by theory and extensive repeated studies, the levels of insight the users gain are a good indication of problem solving performance and effectiveness of the system design (Endsley & Jones, 2011). The second construct, namely *perceived value of analytics outcome*, is a good indicator for the domain value of the analytic outcome, taking away the unpredictability and noises in the real world. It also implies the desirability of which the analytic outcome to be actually used to inform decisions. The third construct, *quality of decision*, provides the most realistic performance of the data analytics systems: the quality of decisions taking into consideration the unpredictability, randomness, and noises in the actual world.

---- This space is intentionally left blank ----

Table 16. Analytical performance at three different stages

Stage	Performance Construct	Description
Intermediate outcome of the analytic process	Extents of insight	This performance is measured from the perspective of knowledge states that the users gain along the analytic process. It measures the comprehensiveness of the analyst's understanding about the problem situation.
Immediate outcome of the analytic process	Perceived Value of Analytic Outcome	This performance is assessed from the perspective of knowledge value. It measures the value of the analytic outcome perceived by the users.
Final outcome of the analytic process	Quality of Decision	This performance is assessed from the perspective of applied decision making, which is the consequence of the actual decision. It measures the quality of the actual decision in the practical settings.

6.2.2 Needs for controlling extraneous variables

The difficulty of evaluating analytical performance is also attributed to the extraneous variables that possibly affects the performance. Besides the effects from the system design, analytical performance is also influenced by the users, the task design, the dataset, and other confounding factors. For instance, different datasets, tasks, and settings naturally will affect the analytical performance differently. To objectively evaluate the impacts of the proposed system, it is important to keep these extraneous factors constant. This is to ensure the changes in analytical performance are due to the proposed system, rather than to the effects of other confounding variables. More importantly, whether or not the performance has been improved or has worsened, it is a relative term that needs to be compared with a baseline.

Based on these requirements, this study has chosen a controlled user study as the data collection method. The user study involves comparing the user analytical performance between the proposed system and a conventional data analytics system. The conventional data analytic system acts as the baseline for the comparison. Here "controlled" means that the design, users, and operating settings were held constant when comparing the analytical performance of the two systems.

To obtain the data needed for the comparison, the user study was followed by a questionnaire survey and interview to ensure data from multiple perspectives of the performance were captured. The user study provides factual data such as time spent, functionalities used, and quantity of reasoning artifacts via system logs. The questionnaire survey is designed to address the qualitative and perceptual aspects of the evaluation. The purpose of the interview is to follow up with the participants for in-depth understanding and reconcile conflicting findings from previous sessions. The qualitative and

quantitative data are converged to evaluate the analytical performance. *Figure 77* shows the processes in data collection and the data type associated with the processes.

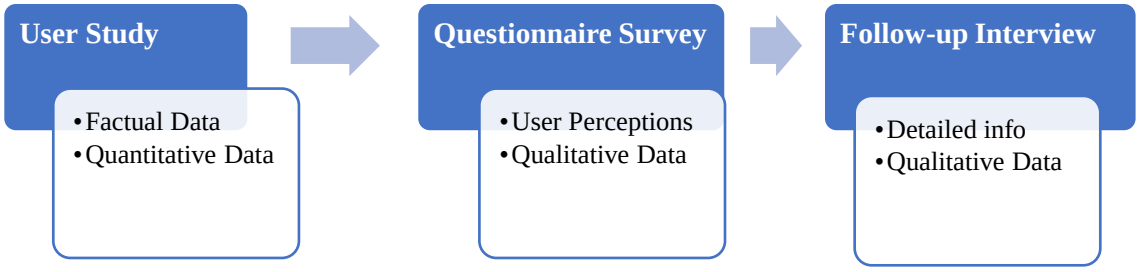


Figure 77. Processes in data collection

6.2.3 Needs for assessing the design as an information system

As an information system, the value of the design also relies on 1) whether the system can be easily learnt and used by potential users and 2) how much effort the users need to exert to use the system.

This study believes it is important to understand how the users perceive the quality of the proposed design as an information system. Particularly in this study, the usability of the proposed system can help to understand the potential trade-offs between costs and benefits of the proposed system, compared to conventional visual analytic systems. For instance, the value of the proposed system is greatly diminished if the costs for learning and using and the system significantly outweigh the benefits it offers. Therefore, the usability of both the proposed system and the conventional system was measured. In this study, usability is measured by usefulness, ease-of-use, learnability, and overall satisfaction.

Besides the usability, the cognitive load is also a critical success factor for information systems, particularly for data analytic systems in which the value is hinged on the user's information processing capacity (Jörn Kohlhammer et al., 2009). The cognitive load theory suggests that information processing happens best when the system design is aligned with human cognition capability. Excessive cognitive loads increase the chances of errors and mistakes, hence affecting the analytical performance of the users. In essence, cognitive load provides insight into how well the proposed system is aligned with user cognition capability and how well it supports the information processing which is the bedrock for the entire data analytic process. This study is interested in examining the cognitive efforts that users need to invest to use the proposed system, compared with using the conventional data analytics systems.

From an information system perspective, demographics data can be used to understand the characteristics of users who can learn and use the system relatively easier than others. From a bigger perspective of this study, demographic variables provide more refined explanations for the analysis results of this study. More importantly, demographic variables are used in this study to examine whether the observed effects were due to the system design rather than to the participants' variability. This

allows the effects of the proposed design to be more reliably evaluated. Two main types of demographics were collected. The first type includes specific demographic variables that related to the stock investment domain, which includes finance knowledge level, stock investment knowledge level, investment strategy preference, and risk profile. The second type is general demographic variables such as education level and occupation type.

6.2.4 Summary of the Evaluation Requirements

Based on the evaluation requirements, three main hypotheses correspond to the analytical performance. Usability and cognitive load also results in several hypotheses. Section 6.3 describes how the constructs, in these hypotheses, such as extents of insights, perceived value of analytic outcome, and usability, were measured. As a result, more specific sets of testable hypotheses are developed in accordance to how the constructs being operationalized. For instance, the construct "extents of insights" is measured by six variables. As a result, six specific *testable hypotheses* were developed. These hypotheses are "testable" because can they be answered directly by statistical tests. Section 6.4 provides a detailed description of how the controlled user study is designed to evaluate the system.

6.3 Operationalization of Constructs

To collect data for the evaluation, the constructs in the hypotheses need to be operationalized into measurable variables. This subsection first discusses the three main constructs for the three different stages of analytical performance, then the two other constructs for usability and cognitive load. For each of the constructs, the discussion is organized in the following way: 1) a brief conceptual overview of the construct, 2) discussions of the variable(s) used to measure the construct, and 3) the resultant testable hypotheses.

6.3.1 Construct 1: Extents of Insight

As described by the HIVE framework (discussed in chapter 4), insights are the knowledge states the users gain at different phases in data analytics task. Insights can be conceptually broken down into 6 progressive levels of insights. Each level refers to a different knowledge state that reflects one's understanding about the analytic problem at hand. By gaining higher-level insights, the analyst can achieve an integrated view of the problem, and deliberately assessed courses of action based on potential future states of the problem. Studies have confirmed that these levels of understanding are keys to effective problem solving (Endsley & Jones, 2011; Haynie, Shepherd, Mosakowski, & Earley, 2010).

Recall that six propositions are derived from the conceptual design framework, as shown in *Table 17* below. These propositions state that the users will have better performance in the problem-solving activities. The six insight components, namely identification insight, perceptive insight, integrative insight, comprehensive insight, predictive insight, and prescriptive insight, are the indicators of the

problem-solving performance of these different activities. In order to evaluate these propositions, the user performance in these problem-solving activities is measured by the insight components.

Table 17. Propositions derived from the conceptual design framework

	Proposition
1	The data analytics systems with capability for ***enabling divergent exploration*** will allow users to perform better in the *search & filter* activity.
2	The data analytics systems with capability for ***enabling managed observations*** and ***enabling exploration convergence*** will allow users to perform better in the *perceive & interpret* activity.
3	The data analytics systems with capability for ***enabling knowledge creation*** will allow users to perform better in the *integrate & synthesize* activity.
4	The data analytics systems with capability for ***enabling assisted situation modelling*** will allow users to perform better in the *connect & build* activity.
5	The data analytics systems with capability for ***enabling predictive reasoning*** will allow users to perform better in the *predict & simulate* activity.
6	The data analytics systems with capability for ***enabling stochastic optimization*** will allow users to perform better in the *optimize & assess risk* activity.

The six insights in this study are derived from the general situation awareness (SA) theory in order to specifically explain user behaviors and cognitive states in the data analytics tasks. The theoretical premise and components of the theory remain unchanged. Thereby, this study asserts that the instruments for measuring situation awareness (SA) are appropriate for measuring its derived counterparts in this study. Moreover, the highly adaptable SA measurement has a long history in the evaluation of various tasks and systems, ranging from complex physical interfaces used in machineries to digital interfaces on jet fighters (Endsley, 1995a; Feng, Teng, & Tan, 2009). SA measurement has gained ever increasing popularity in information system evaluation.

The *situation awareness global assessment technique* (SAGAT) is one of the most well-established situation awareness measurements. This study chose SAGAT as the measurement instrument for two main reasons. Firstly, SAGAT adopts are information processing perspective to assess the extent to which the users can internalize the information resulting from the interactions between the user and the system. This perspective matches the nature of the data analytics systems. Secondly, SAGAT directly measures different awareness levels which can be clearly mapped back to the main components in the situation awareness theory. The separate measurement of each awareness level is aligned with the interests of this study to find out how well the proposed system improves the six different levels of insight over the conventional data analytic systems.

SAGAT provides a measurement framework which requires researchers to develop their task-specific items based on the components in the framework. SAGAT has been found to be able to reliably predict the decision and performance of the users (Endsley et al., 1998). Australia's Department of Defense has also adopted SAGAT to test the relationship between SA and decision making (Stanners & French, 2005). This study adapts the SAGAT's measurement items specifically for the stock investment task. Moreover, this study will implement SAGAT as a post-trial self-rating data collection method. In other words, the SAGAT questionnaire was administered after the user study session.

The primary advantages of the post-trial self-rating approach are its ease of application and its non-intrusive nature. On the other hand, the technique has been criticized for 1) delay between the task and the recall, 2) respondents associate their SA response with their actual performance. This study believes that the delay is reasonably short in this study; therefore, the memory decay effect is negligible and its benefits outweigh the costs introduced by interferences in the freeze probe approach, where the participants are stopped at specific points of the task to answer the questions. In order to prevent the respondents from using their actual performance to reflect on their situation awareness, the actual performance (i.e. the returns of investments) will be withheld from them until the questionnaire survey is completed. *Table 18* summarizes the measurement instruments for measuring the level of insights.

Table 18. An overview of the measurement for extents of insights

Attribute	Description
Measurement instruments	Situation awareness global assessment technique (SAGAT)
Task-specific questions	Yes
Question style	Likert-Scale Questions
Administration style	Self-rating
Timing of administration	Post-trial and single time

There are six measurable variables for the "levels of insight" construct, namely identification insight, perceptive insight, integrative insight, comprehensive insight, predictive insight, and prescriptive insight. The following paragraphs presents the operationalization of these variables into measurement items. The operationalization follows the SAGAT approach in which the dynamic information elements associated with the major processes are identified. Then, these task-specific information elements were used to formulate items that correspond to each level of the insight. *Table 19* shows the measurement items.

Table 19. Variables for levels of insight

Variable	Measurement Items
Identification Insight	To what extent have you adequately identified the factors that are relevant to your stock investment task? To what extent have you adequately identified the stocks that are potentially profitable from the stock market?
Perceptive Insight	To what extent have you adequately understood the factors that are relevant to the stock market? To what extent have you sufficiently understood the stocks based on the how the stocks qualify a combination of the relevant factors?
Integrative Insight	To what extent were you able to combine technical analyses into important knowledge about the stock market? To what extent were you able to incorporate your judgements and assumptions into the understandings of the stock market?
Comprehensive Insight	To what extent have you sufficiently understood the interactions between the factors in the stock market? To what extent have you adequately comprehended the effects of the interactions had on the prices of the stocks?
Predictive Insight	To what extent were you able to forecast the future price trend of the selected stock based on their current price trend? To what extent were you able to forecast future price trend of the selected stock by considering the future movements of other factors in the market?
Prescriptive Insight	To what extent have you adequately evaluated the potential impacts of the stocks to be invested on your earning? To what extent have you sufficiently assessed the potential impacts of the quantity of the stocks to be invested on your earning?

Based on this operationalization, there is one testable hypothesis at the latent construct level and six testable hypotheses at the dimension level. Note that the six testable hypotheses are formulated to

evaluate the six proportions derived from the design framework. The following are the testable hypotheses.

H1: The participants will gain a *higher extent of overall insight* by using the proposed system than using the alternative system

H1a: The participants will gain a *higher extent of identification insight* by using the proposed system than using the alternative system

H1b: The participants will gain a *higher extent of perceptive* insight by using the proposed system than using the alternative system

H1c: The participants will gain a *higher extent of integrative insight* by using the proposed system than using the alternative system

H1d: The participants will gain a *higher extent of comprehensive insight* by using the proposed system than using the alternative system

H1e: The participants will gain a *higher extent of predictive insight* by using the proposed system than using the alternative system

H1f: The participants will gain a *higher extent of prescriptive insight* by using the proposed system than using the alternative system

6.3.2 Construct 2: Value of Analytic Outcome

This second construct measures the perceived value of the analytics outcomes. The higher the value, the more desirable is the analysis outcome in the user's problem-solving context. Studies from knowledge discovery and data mining commonly refer to this value as the "knowledge actionability". An analysis outcome with a higher value also have a higher chance to be actually used to support decision making and to be deployed into the physical world (Cao, 2012).

Table 20. An overview of measurement for Value of Analytic Outcome

Attribute	Description
Measurement instruments	Knowledge Actionability
Task-specific questions	No
Question style	Likert-scale Questions
Administration style	Self-rating
Timing of administration	Post-trial and single time

The quality of the analytic result is an aggregate construct which has multiple dimensions. This study adopts the dimensions from Hui (2014). *Table 21* shows the dimensions and their description.

Table 21. Dimensions for perceived value of analysis outcomes

Dimensions	Descriptions
Understandability	• The extent to which the analytics results can be interpreted and understood. Item: *To what extent do you think the analytic results can be understood in the context of the task?*
Strength	• The extent to which the analytics results is supported by factual data, statistical indices, or other objective indicators. Item: *To what extent do you think the analytic results are supported by factual data and systematic techniques?*
Novelty	• The extent of the analytic results or its elements are new to you Item: *To what extent do you think that the analytic results or its contents are new to you?*
Uniqueness	• The extent of the analytic results is unique and were not easy to be produced by others Item: *To what extent do you think that the analytic results were not easy to be imitated by others*?
Unexpectedness	• The extent to which the analytics result is different from prior analytical experience and deviated from the normal expectation on the result. Item: *To what extent do you think the analytic results are different from your expectations?*
Robustness	• The extent to which the analytics results are consistent despite slight changes in the data or underlying assumptions Item: *To what extent do you think the analytic results are robust to uncertainties?*
Realism	• The extent to which the analytics results are derived from representative models that reflect the condition and constraints of the real-world. Item: *To what extent do you think the analytic results are based on realistic conditions and constraints?*
Comprehensiveness	• The extent to which the analytics results involve sophisticated analytics process or large amount data in a synergic way. Item: *To what extent do you think the analytic results are derived based on in-depth analysis processes?*
Assurance	• The extent to which user perceives the analytic results are likely to succeed Item: *To what extent do you think the analytic results are likely to help to solve the problem successfully?*

Knowledge Building	• The extent to which users can learn from the insight Item: *To what extent do you think the analytic results are able to enrich your knowledge about the task?*
Potential Value	• The extent to which the analysis outcome able to discover opportunities and reduce threat Item: *To what extent do you think the analytic results are able help you in identify opportunities and avoid threats?*
Applicability to Decision	• the extent to which the analysis outcome can be used to support the decision making Item: *To what extent do you think the analytic outcomes can directly provide inputs to the decision making?*

Based on the operationalization, there is one testable hypothesis at the latent construct level and twelve testable hypotheses at the dimension level. The followings are the hypotheses.

H2: The participants will generate analysis outcomes with higher ***overall value*** by using the proposed system rather than the alternative system

H2a: The participants will generate analysis outcomes with higher ***understandability*** by using the proposed system rather than the alternative system

H2b: The participants will generate analysis outcomes with higher ***strength*** by using the proposed system rather than the alternative system

H2c: The participants will generate analysis outcomes with higher ***novelty*** by using the proposed system rather than the alternative system

H2d: The participants will generate analysis outcomes with higher ***uniqueness*** by using the proposed system rather than the alternative system

H2e: The participants will generate analysis outcomes with higher ***unexpectedness*** by using the proposed system rather than the alternative system

H2f: The participants will generate analysis outcomes with higher ***robustness*** by using the proposed system rather than the alternative system

H2g: The participants will generate analysis outcomes with higher ***realism*** by using the proposed system rather than the alternative system

H2h: The participants will generate analysis outcomes with higher ***comprehensiveness*** by using the proposed system rather than the alternative system

H2i: The participants will generate analysis outcomes with higher ***assurance*** by using the proposed system rather than the alternative system

H2j: The participants will generate analysis outcomes with higher ***knowledge building value*** by using the proposed system rather than the alternative system

H2k: The participants will generate analysis outcomes with higher ***potential value*** by using the proposed system rather than the alternative system

H2l: The participants will generate analysis outcomes with higher ***applicability to decision*** by using the proposed system rather than the alternative system

6.3.3 Construct 3: Decision Performance

In this study, the decision performance is assessed by the consequence of the actual decision. The decision performance of the participants is measured from both quantitative and qualitative aspects. The quantitative aspect of the decision is measured by the earning resulted by participants' decisions. In contrast, the qualitative aspect of the decision is measured by how well the participants' decision compared to experts' decision. By taking both perspectives into consideration, the decision performance of participants can be understood from a more complete point of view.

Table 22. Summary of the measurement for decision performance

Attribute	Description
Variable style	Both Quantitative and Qualitative
Administration style	Collected automatically by the systems
Timing of administration	At the end of the task

The quantitative aspect of the decision performance is primarily measured by the earnings generated from their stock portfolio. The total earning is the sum of returns from all the individual stocks within a participant's portfolio. When the user completes the experiment, the total earning will be calculated by the systems based on the stock prices at the end of the investment period. Although the total earnings can be objectively measured and compared, this study believes that it should not be used as the sole measurement for the decision performance. This study also measured the participants' *earning above random baseline* (EARB). Random baseline earning is the average earning that a participant will earn by chance by simply investing in random stocks. Earning above random baseline is the result of total earning subtracted from the random baseline earning. It is important to examine the earning above random baseline to understand the effect size of the proposed tool compared to a random chance. If the effect of the proposed system is not better off than investing in random stocks, the usefulness of the proposed system would be greatly diminished.

The qualitative aspect of the decision performance is measured by how well the participants' decision compared to the experts' decision. Three experts from the stock market field were recruited to select 5 stocks from both of datasets. The experts chose the stocks based on their domain knowledge, which included judging from the stock price, financial health of the companies, and other quality of the stocks. Repeated selections of the same stock will be considered as one. The exercise resulted in two lists of 9 stocks and 7 stocks, respectively, for the two datasets, in which each dataset was used by half of the treatment and the control sessions. Then each of the participants' selections were compared against the experts' selection of the same datasets and the percentage of match was calculated. For example, if a participant has 3 of 5 selections matched the experts' selection that is a 60% of match. Due to randomness and imperfect relationships in the actual price movement, the quantitative measure of the decision may not represent the true effects of the proposed system. The qualitative aspect provides a measurement that alleviate this issue. *Table 23* summarizes the indicators of decision performance.

Table 23. Variables and their measurement for decision performance

Variable	Description
Total Earning	• **Quantitative aspect** of decision performance • Represents the main goal of stock investors • Measured by the total earning (or total loss) of the entire portfolio
Earning above random baseline	• **Quantitative aspect** of decision performance • Represents the overall quality and balance of the portfolio • Measured by the total earning minus random earnings
Percentage match against experts	• **Qualitative aspect** of decision performance • Represents the subjective quality of the stocks in the portfolio • Measured by the percentage of match against experts' choices.

Statistical Hypothesis (H3):

H3a: The participants will generate ***larger total earnings*** by using the proposed system rather than the alternative system

H3b: The participants will generate ***larger earning above random baseline*** by using the proposed system rather than the alternative system

H3c: The participants will have ***higher percentage match against experts*** by using the proposed system rather than the alternative system

6.3.4 Construct 4: Usability

The usability of the proposed system as an information system refers to the extent to which it can be used by the users to achieve their objective, with effectiveness, efficiency, and satisfaction. In this study, usability can also help to understand the potential trade-offs between the costs and benefits of the proposed system. The costs can be the learning-curve and extra efforts to use the system, while the benefits are the resultant performance improvements and user's needs being met. *Table 24* provides an overview about the measurement of usability in this study.

Table 24. An overview of measurement for usability

Attribute	Description
measurement instruments	Usability (Lund, 2001),
Task-specific questions	No. General questions.
Question style	Qualitative Assessment
Administration style	Self-rating
Timing of administration	Post-trial and single time

This study adopts Lund's (2011) notion of usability. Usability is measured by four variables, namely user satisfaction, usefulness, ease-to-use, and learnability. This study added *continuous usage intention*, as the eventual success of the system largely relies on the willingness of users to continuously use the system in the future (Tsai, Chien, & Tsai, 2014). *Table 25* shows the variables, their description, and items.

Table 25. Variables and measurements for usability

Variables	Description
User Satisfaction	User satisfaction measures the overall attitude of the user toward the system. It has been common agreed to be a key success factor of a computer software. In this study, this variable provides a yardstick for comparing the overall satisfaction between the prototype and the conventional data analytic system. Item: Overall, I am satisfied with the software.
Usefulness	Usefulness measures to the degree to which the users believe that the system is meeting their needs. Usefulness has been well-established as one of the most influential factors for user satisfaction (Calisir & Calisir, 2004). Usefulness of

	the prototype is a critical factor to assess whether the extra analytic supports are actually delivering added value to the users. Item: The software is useful for supporting me to accomplish the task.
Ease-to-use	Ease-to-use measure the degree to which a person believes that the system is simple to use. In this study, ease of use of the prototype system is a key factor to be assessed as it is commonly found to affect the overall satisfaction and the system continuance. If the users find that an analytic tool is too difficult to use, it will jeopardize the tool's practicality. Item: The software is easy to use.
Ease of Learning	Ease of learning refers to the degree to which a person believes that the system is easy to learn. Ease of learning has been commonly found to strongly contribute to overall user satisfaction (Calisir & Calisir, 2004). In this study, it is important to evaluate the learnability of the prototype system because it helps this study to understand the extent of effort required by the participants to learn the prototype, particularly compared to conventional visual analytic systems. Item: The software is easy to learn.
Continuance Intention	Continuance intention to use measures the tendency to which a user to continuously use the system in the future. For an analytic system to be valuable, it is critical to assess the user's intention to continuously use the system in the long run; that is, beyond the initial acceptance and satisfaction of the system. Item: I would like to continue using this system in the future.

Statistical Hypothesis (H4):

H4a: The participants will perceive the proposed system to be more satisfying to use than the alternative system

H4b: The participants will generate ***larger earning above random baseline*** by using the proposed system rather than the alternative system

H4c: The participants will have ***higher percentage match against experts*** by using the proposed system rather than the alternative system

6.3.5 Construct 5: Cognitive Load

Cognitive load theory is concerned with the knowledge building in complex cognitive tasks, in which the knowledge building is often overwhelmed by the number of interactive information elements that need to be processed simultaneously before meaningful knowledge building can commence (Paas, Van Gog, & Sweller, 2010). This description fits nicely to data analytics. For the users interacting with interfaces, data, and models in data analytic, the cognitive load is a critical factor for the information processing performance. Excessive levels of cognitive load increase the chances of errors and mistakes, thus impairing the user analytical performance.

According to cognitive load theory, during the complex cognitive task, three types of cognitive load can be imposed on the user's cognitive resource. They are 1) *intrinsic load*, which determined by the task complexity and its information, 2) *germane load*, that is the user's effort contributing to the meaningful knowledge building, and 3) *extraneous load*, which was induced by unclear features or interfaces but did not contribute to the meaningful knowledge building. The goal of information system design is to increase germane load, while reducing extraneous load.

Cognitive load can be measured through self-rating Likert-scale items, by which the users need to rate the amount of mental effort they have invested in completing the task. Paas, Tuovinen, Tabbers, and Van Gerven (2003) stated that mental effort rating scales have proved to be valid, reliable, and unobtrusive. Recent validation of the instrument confirmed its reliability and validity (Leppink, Paas, Van der Vleuten, Van Gog, & Van Merriënboer, 2013). This study adopts their latest ten-item questions for measuring cognitive load in this study. *Table 26* shows the variables and their measurement items.

Table 26 Variables of Cognitive Load

Variables	Description
Intrinsic Load	Intrinsic load is associated with the inner nature of the task, which is largely influenced by both the number of information elements and the interactivity of those element. It refers to the cognitive resource needed to store and process the information elements. Questionnaire Items: • The data covered in the task were very complex • The task covered data that I perceived as very complex • The task covered concepts that I perceived as very complex
Germane Load	Germane load is associated with the mental efforts that are directly relevant to knowledge building activities, such as constructing and refining mental schemata from the information elements (Vandewaetere & Clarebout, 2013). Questionnaire Items: • The task really enhanced my understanding of the data covered in the task • The task really enhanced my understanding of the stock market in the task • The task really enhanced my understanding of the stock market analysis • The task really enhanced my understanding of the concepts in stock investment
Extraneous Load	Extraneous cognitive load refers to the load induced by the design of the system. The load can be influenced by how the information is presented and how the user interactions is designed. Extraneous load is considered to be the cognitive resources that did not contribute knowledge building activities. Questionnaire Items: • The system used during the task was very unclear • The system was, in terms of learning, very ineffective • The system was full of unclear design or interface

Statistical Hypothesis H5:

H5a: The participants will perceive a *lower extent of intrinsic load* by using the proposed system than using the alternative system

H5b: The participants will perceive a *lower extent of germane load* by using the proposed system than using the alternative system

H5c: The participants will perceive a *lower extent of extraneous load* by using the proposed system than using the alternative system

6.4 Designing the User Study

This section presents the user study, which provides a contrived setting for data collection. Details such as user study participants, procedure, tasks and dataset used are presented.

6.4.1 Participants

The user study requires the participants to analyze a stock market and develop a stock portfolio. This task requires the participants to have the basic financial knowledge, such as the ability to coarsely understand the components in a balance sheet report and commonly used financial ratios. However, the participants are not required to have actual experience in stock investment. This is because the emphasis of this study is the general problem-solving process. The stock investment task is intended to emulate a complex analytic problem. Novice users who have minimal experience in stock investment are suitable for the user study because the task can better represent a complex problem for these users. In contrast, professional traders were not recruited because their familiarity with the stock market can help them to "cheat" in the task. The user study used actual historical stock market data from the United States. Professional traders might be able to make profitable decision by recalling the historical performance of the stocks, rather than based on the data analytics.

Given this requirement, the target participants of this study are undergraduate students from QUT business school. A more homogeneous source of the participants would reduce the confounding factors induced by participants. For examples, as undergraduate students from the same school, they are likely to 1) possesses a similar level of business-related knowledge, 2) fall within the same age group, and 3) have basic financial knowledge. Ethical clearance has been obtained from QUT Office of Research Ethics and Integrity to ensure that all the procedures used in this study comply with required standards. The approved recruit email and flyer templates from QUT Ethics were used to inform the students about the user study. Refer to Appendix B for the ethical clearance approval. Before students signed up for the user study, they were informed about the task, the duration required, and the compensation.

Given time restraints and difficulties in recruiting participants for the user study, this study aimed to recruit a minimal number of 30 participants. The number allows differential statistic tests such as T-tests, ANOVA, and Mann-Whitney test to produce results with sufficient reliability. Using 30 participants would approximately produce a margin error of +/- 7% at a confidence level of 90% in this study. There were 38 participants were recruited for the user study. Four withdrew before completing the user study and four participated in a pretest session. This resulted in 30 participants completing the user study.

6.4.2 Activities in the User Study

The primary purpose of the user study here is to enable data collection by creating a contrived setting in which the researcher can manipulate the variables of interest whilst controlling confounding variables such as the users, dataset, analytical task, computer hardware, and time of the day.

In order to evaluate the design effects of the proposed system, the proposed system and the conventional data analytics system were used in the two separate sessions for comparison. This results in two sessions, namely treatment session and control session. The controls session allows this study to measure the natural variability of the dependent variable, to provide a means of measuring error in the experiment, and also to provide a baseline to measure against the proposed system (Carpi & Egger, 2008). Each participant was required to participate in both the treatment and the control sessions. Each session comprises the processes shown in *Figure 78*.

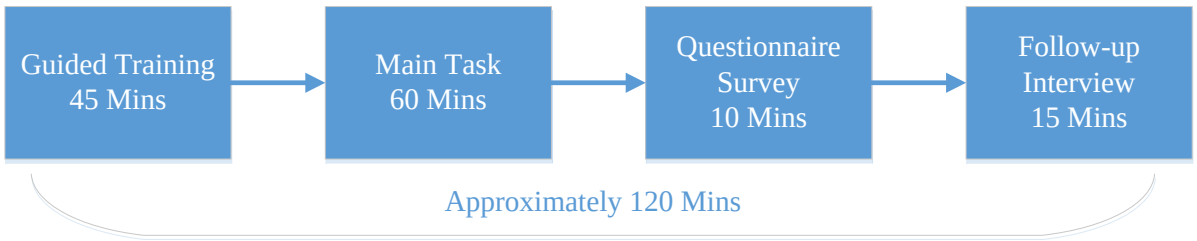


Figure 78. Activities in each session of user study

Each session was approximately 150 minutes. In each of the sessions, the participants first went through 45 minutes of guided training on the system. The training was meant to teach the participants about the functionalities and interaction with the software systems; they would not be taught about how to make investments. A standardized training protocol was used to ensure that every participant received the same information and the same extent of training. In the next 60 minutes, they would use the system to analyze a stock market and make their investment decision. The dataset used in the training and practice were different from the datasets being used in the actual session. Next, the participants were required to answer a questionnaire survey which takes approximately 10 minutes. The participant would then participate in a 15-minute one-on-one interview. Extra time was taken into consideration for filling the consent form, debriefing, and other unforeseeable delays.

To minimize the confounding efforts from the user study settings, this study ensures that the same participant used the same computer setup and the same room for the two sessions, at roughly the same time of the day. Due to the long and taxing activities in each session of the study, each participant was required to undertake the two sessions on two separate days. This is to ensure the participants have a fresh mind and to perform optimally in both sessions.

6.4.3 Task Design

The 60-minute main task in the user study requires the participants to analyze a stock market and make an investment decision to maximize the return. Each participant was allocated USD $500,000 as the virtual investment capital. There are 50 stocks in a stock market. The participants were required to invest the capital in a portfolio with a maximum number of 5 stocks. The datasets available to the participants contains 4.5 years of historical data, dated from January 2010 to June 2015. The participants were to make the investment decision at of 30 June 2015 and the returns were calculated based on results on the last of trading day of December 2015. The investment returns of each participant were independent and were not influenced by the investment returns of other participants. The following bullet points summarize the task:

- Scenario (Same for Session 1 & 2)
    - Domain: Stock market analysis
    - Stock Market: 50 Stocks from S&P 500
    - Point in Time: 1st July 2016
    - Historical Data: January 2010 – June 2015

- Objective
    - To **maximize** the returns of investment at the end of December 2015

- Constraints
    - Capital Available: **USD $500,000**
    - Number of Stocks to Invest: **5 Stocks at Max**

In order to provide an incentive for the participants to carry out the tasks with the level of effort a reasonable person would take on making a financial decision, the participants who had the best decision performance were awarded with $300 worth of vouchers. The participants were informed about the incentive when they were recruited and right before they started the user study. All participants were also compensated for their participations in the user study.

Two datasets were used in the main tasks. The data included the historical price of the stocks, company-specific information, industrial-specific trend, and macroeconomic factors. *Figure 79* shows the data categories in the datasets.

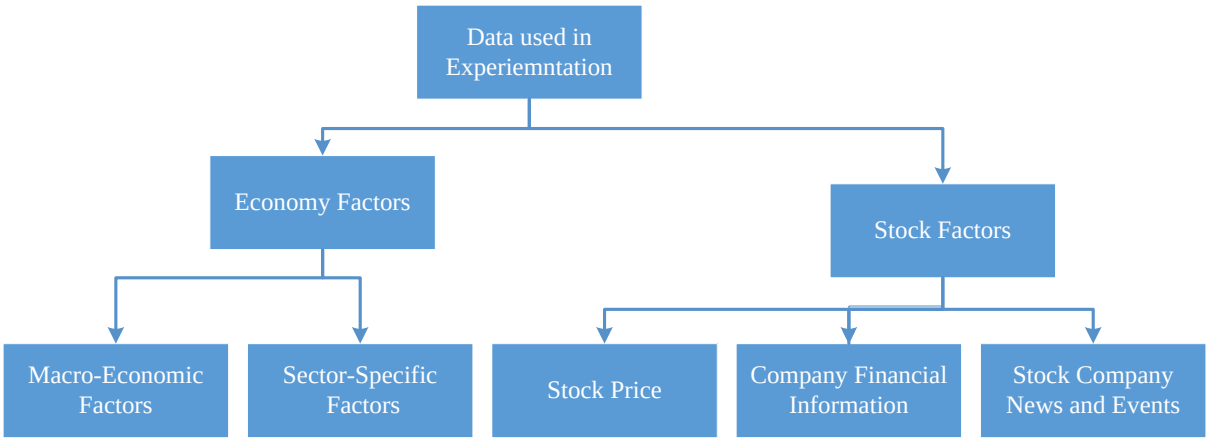


Figure 79. Data types included in the user study's datasets

The stocks used in the user study are the subset of S&P 500 index. The selection of the stocks from the 500 stocks requires a rigorous approach to avoid systematic bias. Firstly, stocks from well-known companies such as Apple Inc., Nike, and Starbucks were removed from the list. The purpose is to avoid the participants to blind guess by choosing companies that are generally known to perform well. Subsequently, the remaining stocks were divided into two groups made up of the top 50 stocks and the bottom 50 stocks based on their price performance (a half-year term). Then, 50% of stocks were randomly selected from each group to create a stock market to be used in the user study. As the result, there are two stock markets, each consisting of 50 stocks, namely dataset A and dataset B used in the main task.

The participants were randomized on the sequence: whether they undertook the treatment session first or the control session first. As a result, half of the participants participated in the treatment session first, while the other half participated in the control session first. The purpose was to normalize the participants who took the treatment group first and those who were involved in the control group first. Moreover, two different datasets were used in the treatment and control sessions. Using the same method, participants were randomized to use datasets A and B. *Figure 80* illustrates the allocations of participants according to the datasets and session sequence.

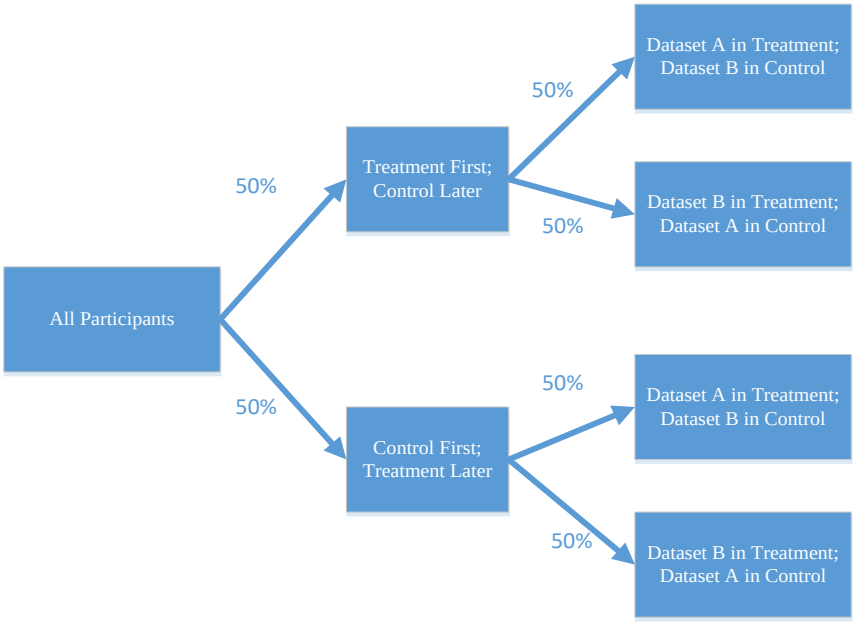


Figure 80. Randomization of the session sequence and datasets

6.4.4 Data Analysis Method

As noted, each participant was required to participate in both the treatment and the control sessions. This arrangement also implies that within-subject analyses were used to analyze the data, as opposed to between-group analysis. Within-subject design is a more stringent evaluation than the between-group design. The within-group effect is a measure of how much an individual participant tends to vary over the different interventions. The main advantages of within-subject design include 1) better generalization capability and 2) robustness to the variability among the participants (Hall, 1998). Analysis techniques for within-subject design automatically offset the variability in the individual factors. In other words, the design also does not require highly homogeneous participants in each group, as required in between-subjects design. Therefore, the analysis results from the heterogeneous pool of participants can be better generalized to the general users. More importantly, the stringent nature of the within-subject techniques allows it to have higher statistical power, compared to between-group techniques with the same sample size. In short, analysis techniques for within-subject design tend to reduce the error variances and thus the results can be interpreted with higher confidence.

6.4.5 Pre-test

Before conducting the user study, a pre-test was conducted for the following purposes:

- To test the facial validity of the measurement instruments
- To ensure the instructions for the user study can be clearly understood

- To ensure the time allocated for the tasks is realistic
- To pretest the questionnaire
- To get a feel of the data

Firstly, the contents of the questionnaire were reviewed by two senior academic staff for the questions' surface validity, sufficiency, and understandability. Facial validity is the degree to which a measurement instrument (e.g. questionnaire items) measures what it is intended to measure (Sekaran & Bougie, 2010). Some of the questions have been reworded for better understandability. Subsequently, four additional participants were recruited to participate in the pre-test. These participants did not participate again in the actual user study. The pre-test's participants were not being told that they were participating in a pre-test or trial run. The same protocol and procedure were used to run the user study. The exception is that the data collected from the pre-test is not included in the data analysis. The data from the pre-test was used to conduct a quick analysis to check for any major flaw in the user study and the questionnaire design. After discussion with the supervisory team, and with no major nor systematic errors found, approval was given by the supervisory team to roll out the actual user study with the remaining of 30 participants.

---- This space is intentionally left blank ----

Chapter 7
Evaluating the Designs

7.1 Overview

This chapter presents and discusses the results from the hypotheses testing. The first section in this chapter briefly recapitulates the three main hypotheses. Section 6.2 presents the participants' demographics. Section 6.3 provides the results of the main hypotheses testing. The results are organized according to the three main hypotheses, which are divided into three subsections within Section 6.3. Then, section 6.4 analyzes the usability and cognitive loads of the systems. Lastly, the results are discussed in detail in section 6.5.

7.2 Demographics of the Participants

30 participants have successfully completed their tasks in the user study, while four withdrew before completing the user study. They are undergraduate students from QUT business school. On average, the participants have spent 91 minutes in the user study's main tasks. The objective of this section is to demonstrate that the 30 participants are realistically distributed across different categories of demographic variables. *Table 27* presents distributions of the participants on various demographic variables. Most of the variables such as ***financial knowledge***, ***investment knowledge***, ***investment strategy***, and ***investment risk profile*** indicate that the samples are mainly distributed around the median categories. This implies that the samples do not bias toward a particular category, and hence the samples can be considered as a realistic representation of the common users from the stock investment domain. The remaining demographic variables, ***investment experience*** and ***area of specialty,*** show that the participants can be divided into two roughly equal categories. Using the investment experience as an example, 53% (16 participants) have no investment experience, while the other 47% (15 participants) have investment experience. Such distribution allows using highly-robust comparison analyses such as T-test to use to compared various variables of interests between the two categories. Such divisions also meaningfully represent different cohorts of potential users in the market. As a result, the findings would provide useful insights into how the proposed system would be perceived by different users in the market, thus allowing this study to gauge the potentials of the proposed system in the competitive landscape of data analytics systems.

Table 27. Demographics of the participants

Demographic Variables	Categories	Frequency	Percentage
Financial knowledge	None	0	0.0
	Minimal	3	10.0
	Basic	7	23.3

	Well	9	30.0
	Very well	9	30.0
	Excellently well	2	6.7
	Total	30	100.0
Investment knowledge	None	0	0.0
	Minimal	8	26.6
	Basic	12	40.0
	Well	7	23.3
	Very Well	3	10.0
	Excellently well	0	0.0
	Total	30	100.0
Investment Experience	Never traded	16	53.3
	Traded a few times	4	13.3
	Trade fewer than 10 transactions a year (active)	6	20.0
	Trade more than 10 transactions a year (active)	3	10.0
	Trade more than 25 transactions a year (active)	1	3.3
	Trade Professionally	0	0.0
	Total	30	100.0
Investment Strategy	Buy and sell on the same day	2	6.7
	Buy and sell within a month	3	10.0
	Buy and sell between 3 to 6 months	7	23.3
	Buy and sell between 6 to 12 months	7	23.3
	Buy and sell between 1 to 5 years	7	23.3
	Buy and sell for longer than 5 years	4	13.3
	Total	30	100.0
Investment Risk Profile	Willing to risk losing more than 100% of the capital	0	0.0
	Willing to risk losing 100% of the capital	1	3.3
	Willing to risk losing 50% of the capital	9	30.0
	Willing to risk losing 25% of the capital	14	46.7
	Not willing to risk losing ANY of the capital	6	20.0
	Total	30	100.0
Area of Specialty	Finance	16	46.7
	Other business areas	14	53.3
	Total	30	100.0

7.3 Evaluating the Main Hypotheses

7.3.1 Evaluating the Effects on the Insight Components

Hypothesis H1 speculates that the participants will be able to gain higher extents of the insights on the analytic problem at hand in the treatment session, compared to the control session. This main hypothesis first compared the overall insight, which is a latent variable between the two sessions. Then each dimension of the latent variable is tested to have a greater understanding of how various aspects of the insight were affected by the proposed system. As noted in Subsection 5.6 titled “operationalization of constructs”, there are three levels of insight and each level contains two dimensions. This resulted in 6

testable hypotheses at the dimension-level. *Figure 81* shows the relationships between 1) the construct, 2) its dimension, and 3) the corresponding tests (denoted by H1, H1a, H1b, and so on).

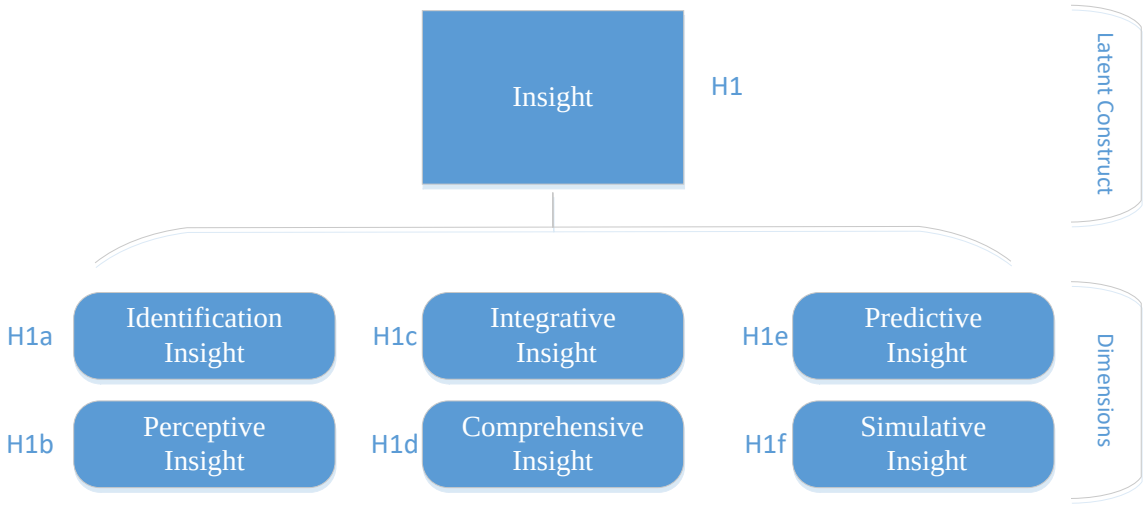


Figure 81. Insight, its dimensions, and the tests

Table 28 shows the paired-sample t-test result of the overall insight between the two sessions. The result indicates the mean difference between the sessions is statistically significant ($p \leq 0.001$). It demonstrates that the participants on average have achieved higher extents of overall insight in the treatment session compared to the control session. According to Cohen's guideline on effect size, the effect size of the difference between two sessions is in the medium-to-large category.

Table 28. Paired-sample t-test on the overall insight

Variable	Experiment Session	Mean	Std. Dev	Mean Difference	Effect Size	Significance
Insight	Control	3.87	0.2612	.2883	0.653	.001***
	Treatment	4.06	0.2458			

A series of follow-up tests at the dimension-level were conducted to examine which aspects of the insight were scored higher than others. Similarly, the paired sample t-test was used for the testing. *Table 29* shows the results of the tests at the dimension-level.

Variable	Experiment Group	Mean	Std. Dev	Mean Difference	Effect Size (Cohen's d)	Significance
Identification Insight	Control	3.93	0.583	.033	0.056	0.758
	Treatment	3.97	0.472			
Perceptive Insight	Control	4.28	0.468	-.200	0.300	0.110
	Treatment	4.08	0.417			
Integrative Insight	Control	3.98	0.549	.167	0.320	0.086*
	Treatment	4.15	0.326			
Comprehensive Insight	Control	3.65	0.589	.317	0.440	0.023**
	Treatment	3.97	0.454			
Predictive Insight	Control	3.73	0.612	.250	0.360	0.057*
	Treatment	3.98	0.404			
Prescriptive Insight	Control	3.83	0.562	.367	0.470	0.016**
	Treatment	4.20	0.610			
*p < 0.10; **p < 0.05						

Table 29. Paired T-tests on the six dimensions of insight

For the **identification insight**, the means between the two experiment sessions are largely indifferent (p-value 0.758). The result indicates that the proposed system did not cause the participants to gain a higher or lower extent of identification insight. This is a control hypothesis. The features to improve identification insight are implemented in both the systems. It is expected that the identification insight gained by the participants should not be significantly different between the two sessions. The intention is to provide a clue as to whether the results of the hypotheses testing are the effects of the features inclusion, rather than other confounding factors.

For the **perceptive insight**, the mean difference between the two experiment sessions is not highly significant, p = 0.11, but it is close to a significance level of 90%. It is noteworthy because the result indicates that there is a less than moderate chance that the participants have gained higher extents of perceptive insight from using the control system. The difference comes with a small effect size, d = 0.30.

For the **integrative insight**, the means between the two sessions show a moderate trend toward a conventional significance level, p = 0.86. The result shows that, on average, the participants tended to gain a higher extent of integrative insight in the treatment session than the control session. The effect size of the mean difference is small, d= 0.32.

For the **comprehensive insight**, the means between the two sessions are statistically different, p = 0.023). The result indicates that the participants gained a significantly higher extent of comprehensive insight in the treatment session, compared to the control session. The effect size of the difference is medium, d = 0.44.

For the **predictive insight**, the mean difference between two experiment sessions was on the verge of conventional significance, p = 0.057. The result indicates that higher extents of predictive insight were gained by the participants in the treatment session than in the control session. The effect size of the difference is small, d = 0.36.

For the **prescriptive insight**, the means difference between the sessions is highly significant, p = 0.016. The result shows that significantly higher extents of prescriptive insight were gained by the participants in the treatment session compared to the control session. The effect size of the difference is medium, d = 0.47.

7.3.2 Evaluating the Effects on the Value of Analysis Outcomes

Hypothesis 2 speculates that the participants will perceive the analysis outcome in the treatment session is of higher value than the control session. This main hypothesis is first tested at the latent construct level, then each dimension of the construct is examined to better understanding how the various dimensions were being perceived different between the two sessions. The latent variable contains 12 dimensions. This resulted in 12 testable hypotheses at the dimension level.

Paired-samples t-tests were used to compare the overall value and its dimensions between two sessions. Following the paired-sample t-test procedure, extreme outliers in the data were identified. Inspection of their values did not reveal them to be extreme and they were kept in the analysis. The assumption of normality was not violated, as assessed through the Q-Q plot of each dimension.

Table 30 shows the result of the main hypothesis testing. The mean difference between the treatment and the control sessions is found to be statistically significant (p-value 0.015). The result indicates that participants perceived higher value from the analysis outcomes in the treatment session compared to the control session. Based on Cohen's guideline on effect size, a medium effect size is found (Cohen's d = 0.46).

Table 30. Paired-sample T-test on value of analytic outcome

Variable	Experiment Session	Mean	Std. Dev	Mean Difference	Effect Size	Significance
Perceived Value of Analytic Outcome	Control	3.84	0.343	.2033	0.460	0.018**
	Treatment	4.04	0.311			

Follow-up tests at the dimension level are run to give greater explanatory power to the result. Similarly, paired t-tests were used for the testing at the dimension level. *Table 31* shows the results.

Table 31. Paired-sample T-test on dimensions of the value of analytic outcome

Variable	Experiment Group	Mean	Std. Dev	Mean Difference	Effect Size	Significance
Understandability	Control	4.27	0.583	-0.333	0.377	0.048**
	Treatment	3.93	0.691			
Strength	Control	4.17	0.648	0.133	0.171	0.354
	Treatment	4.30	0.651			
Adaptability	Control	4.00	0.788	0.300	0.304	0.107
	Treatment	4.30	0.750			
Uniqueness	Control	3.13	0.571	0.367	0.395	0.039**
	Treatment	3.50	0.938			
Unexpectedness	Control	2.87	0.730	0.167	0.200	0.283
	Treatment	3.03	0.718			
Robustness	Control	3.73	0.980	0.367	0.334	0.078*
	Treatment	4.10	0.803			
Realism	Control	4.27	0.944	-0.200	0.183	0.326
	Treatment	4.07	0.583			
Comprehensiveness	Control	4.17	0.592	0.167	0.211	0.258
	Treatment	4.33	0.606			
Assurance	Control	3.80	0.961	-0.300	0.284	0.130
	Treatment	3.50	0.777			
Knowledge Building	Control	4.03	0.615	0.333	0.378	0.048**
	Treatment	4.33	0.711			
Potential Value	Control	4.03	0.490	0.200	0.262	0.161
	Treatment	4.17	0.531			
Applicability to Decision	Control	4.10	0.712	0.333	0.361	0.057*
	Treatment	4.43	0.568			

Understandability. The *understandability* to which the participants perceived in the control session was significantly higher compared to the treatment session, p = 0.048. The effect size is marginally close a medium-size effect, d = 0.377.

Strength. The mean difference of analysis outcomes' *strength* between the treatment and control sessions was not statistically significant, p-value 0.354. The result indicates that the strength of the analysis outcomes from both the sessions was not perceived to be different by the participants.

Adaptability. The mean difference of the *adaptability* between the treatment and control sessions was on the verge at 90% significance level with a p-value of 0.107. The result indicates that there is a less than moderate chance that the participants had perceived that the analysis outcomes from the treatment session is higher in term of adaptability.

Uniqueness. The mean difference of uniqueness between the treatment and control sessions was statistically significant, p = 0.039. The result shows that the uniqueness of the analysis outcomes from the treatment session was scored significantly higher by the participants. The effect size of the difference is close to medium effect, d = 0.395.

Unexpectedness. The mean of strength between the treatment and control sessions was not significantly different, p = 0.283. The result indicates that the unexpectedness of the analysis outcomes from both the sessions was not perceived to be different by the participants.

Robustness. The mean difference of robustness between the treatment and control sessions was marginally significant, p = 0.78. The result suggests that the participants tended to perceive the analysis outcomes from the treatment session are more robust than the control session. The effect size is small, d = 0.334.

Realism. The mean difference of robustness between the treatment and control session was not statistically significant, p = 0.326. The result indicates that the realism of the analysis outcomes from both the sessions was not perceived to be different by the participants.

Comprehensiveness. The mean difference of comprehensiveness between the treatment and control session was not statistically significant, p- = 0.258. The result indicates that the comprehensiveness of the analysis outcomes from both the sessions was not perceived to be different by the participants.

Assurance. The mean difference of Assurance between the treatment and control sessions was not statistically significant, p = 0.130. The result indicates that the analysis outcomes from both the sessions was not perceived to be different by the participants, in term of its likeliness to successfully solve the analytic problem.

Knowledge building. The mean difference of knowledge building the between treatment and the control sessions was partially significant, p = 0.67. The result shows that the participants tended to perceive the analysis outcomes from the treatment session contains higher value for knowledge building. The difference has a small-to-medium effect size.

Strategic value. The mean difference of the analysis outcomes' strategic value between the treatment and the control sessions is found to be significant, p = 0.161. The result indicates that the strategic value of the analysis outcomes from both the sessions was not perceived to be different by the participants.

Applicability for Decision Making. The mean difference of this dimension between the treatment and control sessions was very close to a conventional significance level of 95% (p-value 0.57). The result indicates that the participants perceived that the analysis from the treatment sessional is more applicable for supporting the decision making. There is a small-to-medium effect size.

7.3.3 Evaluating the Effects on Decision Performance

Hypothesis H3 theorizes that the participants' decision performance will be higher in the treatment session compared to the control session. This hypothesis is evaluated from both quantitative and qualitative aspects of the participants' decision. Firstly, as a quantitative aspect of the decision, the earnings of the participants between the two sessions were compared. Secondly, as the qualitative aspect of the decision, the stocks selected by the participants' as investment choices were compared against experts' choices.

To examine the quantitative aspect of the decisions, the total earnings of the participants from the two sessions were compared. Paired sample T-test was used to determine whether the difference in the earnings is statistically significantly. *Table 32* show the T-test result on the total earnings.

Table 32. Paired-sample T-test on total earnings

Variable	Experiment Session	Mean	Std. Dev	Mean Difference	Effect Size	Significance
Total Earnings	Control	-2927.85	29853	6888.98	0.083	0.651
	Treatment	-9816.84	74715			

The result shows no statistically mean difference between the total earnings from the two sessions, p-value 0.651. Nonetheless, a general trend can be observed when comparing the proportion of participants who have total positive earning between the treatment and the control session. The proportion of participants with positive earning has increased from 33% to 50% as shown in Figure 82. A related-samples McNemar test shows the proportional difference between the two sessions was not statistically significant (p-value 0.143).

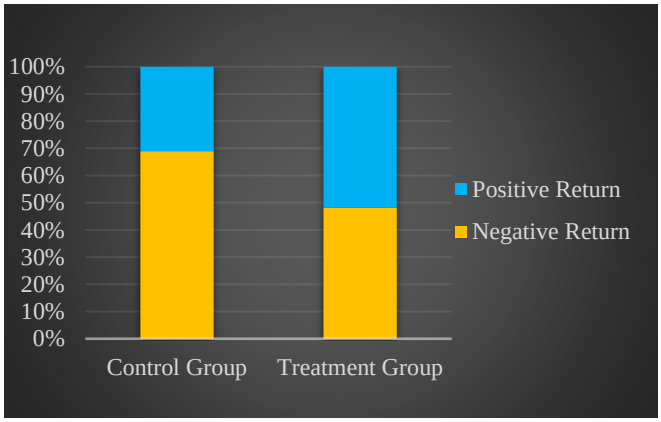


Figure 82. Proportion of participants with positive and negative returns

In addition to the total earning, this study also measured the participants' *earning above random baseline* (EARB). Random baseline earning is the average earning that a participant will earn by chance by simply investing in random stocks. Earning above random baseline is the result of total earning subtracted away the random baseline earning. It is important to examine the earning above random baseline to understand the effect size of the proposed tool compared to a random chance. If the effect of the proposed system is not better off than investing in random stocks, the usefulness of the proposed system would be greatly diminished. *Table 33* shows the paired sample t-test result on the earnings above random baseline between the two sessions.

Table 33. Paired-sample T-test of the earnings above the random baseline

Variable	Experiment Session	Mean	Std. Dev	Mean Difference	Effect Size	Significance
Earning above random baseline	Control	35324.94	29186	6888.98	0.085	0.651
	Treatment	28435.96	74109			

Similar to total earnings, the result shows the difference between the two sessions is not statistically significant, p = 0.651. This study is also interested to examine whether the proportions of participants who have earned above the random baseline in the treatment and control sessions are different. As shown in the following *Figure 83*, the results show the participants who earned above the random baseline have dropped from 90% in the control session to 80% percent in the treatment session. Nevertheless, related-samples McNemar test showed the proportion difference between two sessions was not statistically significant, p = 0.727. The combination of the results shows there is lack of evidence to prove that the quantitative aspect of the participants' decision in the two sessions are different.

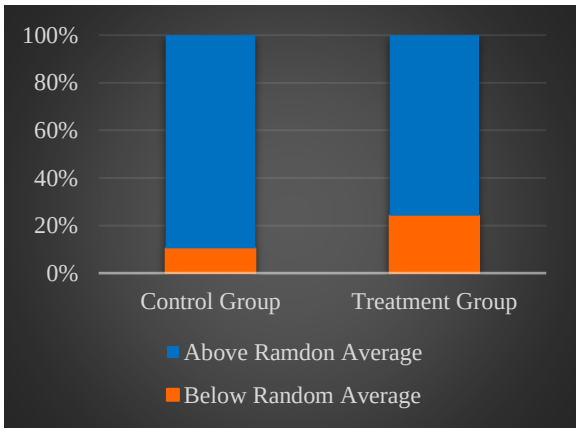


Figure 83. Proportion of participants above and below the random average

In terms of the qualitative aspect of the participants' decisions, the following test compares the proportions of the stock selections of the participant that matched the selections from experts between the two experiment sessions. Three experts from the stock market field were recruited to select 5 stocks

from both of datasets. The repeated selections of the same stock will be considered as one. The exercise resulted in two lists of 9 stocks and 7 stocks, respectively, for the two datasets in which each dataset was used by half of the treatment and the control sessions. Then each of the participants' selections were compared against the experts' selection of the same datasets and the percentage of match was calculated. For example, if a participant has 3 of 5 selections matching the experts' selection, this is a 60% match. Descriptive statistic indicates that, on average, participants from the treatment sessions have higher match percentage of 52.9%, compared to 40.5% from the control sessions. A paired t-test shows that the match percentages between the treatment and control sessions are statistically different (p-value 0.48). The result indicates that difference has a small-to-medium effect size (Cohen's d 0.38).

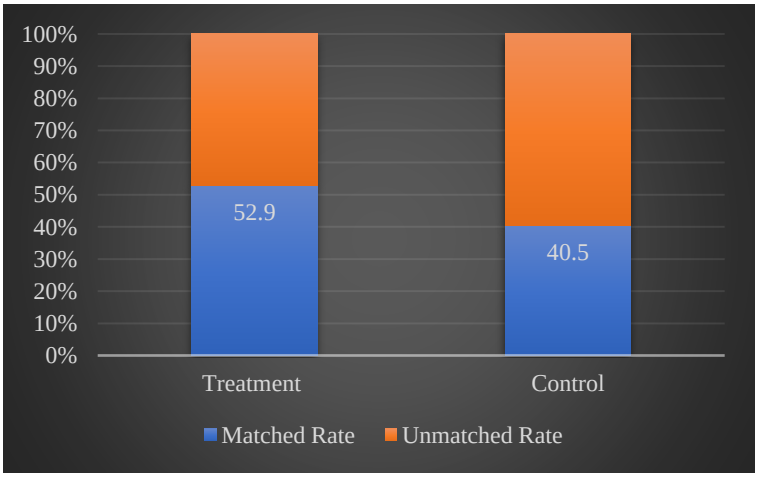


Figure 84. Match percentage between the treatment and control group

From the qualitative aspect of the decision performance, the participants were found to have higher quality decisions in the treatment session. In contrast, the test on the quantitative aspect of the decision performance shows the otherwise. Given the mixed result, the hypothesis testing is inconclusive. Detail discussion will be provided in the discussion section.

7.3.4 Summary of Hypotheses Testing

This section summarizes the results of the hypotheses testing.

Table 34. Summary of the main hypothesis testing

Hypothesis	Description	Result
H1	The participants will gain a higher extent of insights by using the proposed system than by using the alternative system	Supported
H2	The participants will generate analysis outcomes with higher value by using the proposed system than by using the alternative system	Supported

H3	The participants will make decisions with higher quality by using the proposed system than by using the alternative system	Inconclusive

7.4 Evaluating the Designs as an Information System

7.4.1 Usability

This subsection presents the effects of the proposed system on the various dimensions of usability. Paired-Sample T-Tests were conducted to examine whether the mean differences between the treatment and the control sessions are statistically significant. Table 35 shows the results.

Variable	Experiment Group	Mean	Std. Dev	Mean Difference	Effect Size (Cohen's D)	Significance
User Satisfaction	Control	4.27	0.450	-.100	0.140	0.448
	Treatment	4.17	0.648			
Usefulness	Control	4.20	0.551	.300	0.358	0.059*
	Treatment	4.50	0.509			
Easy to Use	Control	4.03	0.669	-1.000	0.923	0.001***
	Treatment	3.03	0.964			
Easy to Learn	Control	4.43	0.568	-1.133	0.947	0.001***
	Treatment	3.30	1.149			
Intention to Use	Control	3.83	0.834	1.67	0.191	0.305
	Treatment	4.00	0.525			

Table 35. Paired sample T-tests on dimensions of usability

User satisfaction. The participants generally perceived a higher level of satisfaction during the control sessions (mean = 4.27) as opposed to treatment sessions (mean = 4.17). However, the result indicates that the difference is not statistically significant, p = 0.448.

Usefulness. The mean difference for the perceived usefulness between the treatment and the control sessions is found to be slightly short at conventional significance level of 95%, p = 0.059. The result indicates that participants have perceived higher level of usefulness during the treatment session than during the control session. The reported effect size is marginally medium.

Easy to Use. The means for the perceived ease of use between the treatment and the control sessions are statistically different, p = 0.001. The participants perceived that the alternative system in the control session is significantly easier to use compared to the proposed system in the treatment session. The reported effect size is large.

Easy to Learn. The means for easy to learn between the treatment and the control sessions are statistically different (p-value 0.001). The participants perceived that the alternative system is significantly easier to learn compared to the proposed system. The reported effect size is large.

Intention to Use. The perceived intention to use from the treatment and the control session are not statistically different, p = 0.305. The result indicates the intentions of the participants to continuously use the two systems were not different.

7.4.2 Cognitive Load

This subsection presents the effects of the proposed system on the various dimensions of mental load. Paired-Sample T-Tests were conducted to examine whether the means of the dimensions are significantly different between the treatment and the control sessions. Table 36 shows the result of the paired-sample t-tests of each dimension.

Variable	Experiment Group	Mean	Std. Dev	Mean Difference	Effect Size (Cohen's D)	Significance
Intrinsic Load	Control	5.27	0.980	0.567	0.732	0.001***
	Treatment	5.83	0.592			
Extraneous Load	Control	4.43	1.194	1.100	1.005	0.001***
	Treatment	5.53	1.042			
Germane Load	Control	5.53	0.937	0.433	0.381	0.045**
	Treatment	5.97	0.999			

Table 36. Paired-sample T-test on the dimensions of cognitive load

Intrinsic Load. The result shows that the mean difference of intrinsic load between the sessions is statistically significant, p = 0.001. This suggests that participants perceived a significantly higher level of intrinsic load during the treatment sessions compared to the control sessions.

Extraneous Load. The result shows that the mean of extraneous load between the sessions is significantly different, p = 0.001. This suggests that the participants perceived significantly higher level of extraneous load during the treatment sessions compared to the control sessions.

Germane Load. The mean difference of the germane load between the sessions are statistically significant, p = 0.045. The result suggests that the participants perceived significantly higher level of the germane load during the treatment session compared to the control sessions.

7.5 Discussion on the Analysis Results

Overall, the findings suggest that the proposed system has the potential to support users to gain higher level of insights during the analysis process. Moreover, the proposed system can help users to derive at analysis outcomes of higher value that more readily inform the decision or to be translated into an action plan. However, mixed results were found in terms of decision performance. There is no decisive evidence to support this study's hypothesis that the proposed system can help users to achieve decision with high quality in the practical settings. The following subsections provide deeper discussions into the analysis results.

7.5.1 Findings pertaining to the level of insights

Key Findings: The participants have gained higher extents of overall insight by using the proposed system compared to the alternative system. At the dimension level, this study found the participants have gained significantly higher extents of integrative insight, comprehensive insight, predictive insight, and prescriptive insight in the treatment session compared to the control session. The identification insight and perceptive insight gained by the participants were not significantly different between the two sessions. *Figure 85* shows the conceptual position of these insights and whether the insights are significantly enhanced by the proposed system. The following sub-subsections 7.5.1.1 to 7.5.1.6 provide detailed discussions on these findings.

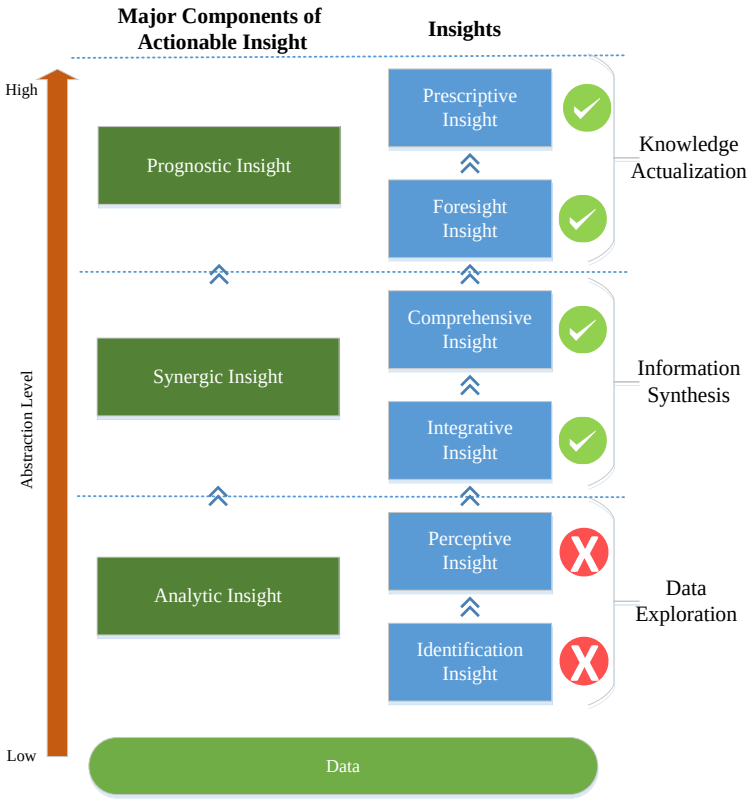


Figure 85. Conceptual position of insights

7.5.1.1 Resulting higher overall insight, mainly attributed to the increases in insights level 1 and 2.

At the overall level, the result indicates participants have achieved significantly higher extents of overall insight from the treatment session compared to the control session. This result implies that the proposed data analytics system would enable users to achieve greater overall understanding of their problem situation, compared to the conventional data analytics system. The proposed system could enhance the users' understanding of the problem situation because the system directly supports users' problem-solving activities in all three phases of data analytics, namely data exploration, information synthesis, and knowledge actualization. In contrast, the conventional data analytics system supports the users only during the data exploration phase, leaving the users to mentally process the information in information synthesis and knowledge actualization phases.

As shown in *Figure 85*, actionable insight comprises the three levels of insight, namely analytic insight, synergic insight, and prognostic insight. The results at the dimension level suggest that the higher overall insight in the treatment session are largely attributed to the significantly largely improved insight level synergic and prognostic insights. This finding indicates that the proposed system built on the design principles can support the users to 1) better understand the problem situation as a whole and 2) better assess the risks of their courses of action in various potential scenarios. However, the proposed system is not able to provide a significant improvement on the users' ability to identify and apprehend the information during the data exploration stage.

Further tests were conducted to understand whether the improved overall insight was due to the usage of the corresponding *advanced modules* meant for supporting the users in the information synthesis and knowledge actualization, but not other confounding factors. Results show there is a medium-strength positive correlation between the proportion of time spent on the advanced modules and the overall insight ($r = 0.489$, $p = 0.006$). Additional tests were run to check whether certain demographics of the participants have caused the variance in the insights gained. The results show all demographics have no significant influence on the insights gained, with exceptional that the participants who prefer the long-term investment strategy tended to score lower on the analytic insight. With the absence of influence from demographics and significant evidence showing positive association between the usages on the advanced modules and the overall insight, these findings suggest that the improvement in overall insights is a result of the usages of the advanced modules in the proposed system.

In short, the overall evidence shows that the proposed system is capable of enabling and scaffolding higher-level interactions between the users and the information which are neglected by conventional data analytics systems. The higher-level human-information interactions such as integrating information, conceptualizing the problem situation, and mentally simulating the outcomes are known to be crucial for solving complex analytic tasks. Such interaction activities bridge the gaps between the low-level data manipulation and higher-level understandings that allow the analysis outcomes to be contextualized in the user's contexts such as goal, constraints, and overall environment. In other words, the higher-level human-information interactions enabled by the proposed system helps the users to gain sufficient understanding of the problem situation to allow them to make informed decision to solve the analytics problem.

7.5.1.2 Resulting higher integrative insight, but the features were underused

The result indicates the participants gained higher extents of integrative insight from the treatment session compared to the control session. This moderately significant result suggests that the proposed system may have the potential to improve users' capability to go beyond dry analysis and turn the analysis into a blend between factual analysis provided by the system and domain knowledge of the users in order to achieve information that is meaningful at problem-solving level. The "enabling knowledge creation" design principle theorised that the users can be more effective in deriving key information that is meaningful at the problem-solving level when they are supported by the integration features 1) to combine findings from multiple low-level analysis into problem-level factors and 2) to integrate quantitative information with subjective knowledge from the users. Data shows there is a medium-strength positive correlation between the usage of the integration features and the participants' level of integrative insight, $r = 0.457$, $p = 0.011$. Furthermore, the gain in the integrative insight is not a function of the participants' demographics. This study, therefore, concludes that the design principle "enabling knowledge creation" can effectively help users to gain better integrative insight.

As the result of the integrative insight is partially significant ($p = 0.086$), the interactions records are further examined to seek out possible causes. Such examination revealed that less than one fifth of the participants have used the integration features. This may suggest that the moderately significant finding was caused by the underuse of the integration feature. In the follow-up interviews, the participants have commented on the integration feature that "*it requires extra efforts and time*", "*not sure how it's going to benefit the analysis*", and "*was not sure what to put in [as the subjective information]*". Given the feedback, the practical value of the integration features will be very limited as long as the users feel that the cost of using the feature outweighs the benefits. Improvement should be done to reduce the interaction costs while allow the high-level knowledge created to have greater uses in a later stage of analysis. This study believes that with more participants using the feature, the effects of design would be more distinguishable and a more decisive result can be established.

In accordance with the evidence, this study concludes that the design principle "higher-level knowledge creation" can be moderately effective for supporting users in the integrate & synthesize activity. It allows the users to be more effective at synthesizing low-level technical analyses into higher-level understandings that are meaningful at the problem-solving level. However, the usefulness of such integration features is conditioned to its interaction cost. The users will be motivated to use the feature more if the interaction required can be simplified and needs less effort. Additionally, this study also believes that the value of the integration feature increases as the size or the abstraction of the analytics problem increases. The abstraction level of the stock market task is very high, and may have rendered the integration feature less important. Nevertheless, this study believes that the "higher-level knowledge creation" design principle holds a promising stand. The integrative features can be especially useful if the analytics problem is highly abstract and the users are experts who have rich domain knowledge that can be incorporated into the analysis. Many studies have stated that analysts must perform information synthesis to gradually build an understanding of concepts or events that are only indirectly supported by the raw information. Moreover, the integration features in this study are one of the very few designs or prototypes that have brought the idea beyond the conceptual stage.

As an additional finding, this study also discovered that the participants' integrative insight has even higher correlation with the number of stocks added into the situation model, $r = 0.511$, $p = 0.004$. The number of stocks in the situation model is a usage indicator of comprehensive feature that is theorized to enhance users' comprehensive insight. Its unexpected effect on the users' integrative insight may suggest that the ability to integrate multiple stocks in a single situation model is also crucial for the users to produce higher level understanding about the problem situation. In other words, this comprehensive feature could have unexpected effects on insight besides the theorized one. The current proposed system allows only up to 5 stocks in a situation model to reduce the model complexity and to avoid a cluttered screen. For better supporting users in deriving integrative insight, future improvements could allow the users to include greater number of stocks in a single situation model. However, each additional stock added to the situation model will exponentially increase the number of factors in the model. Hence, the challenges that remain are to find a representation method that can reduce the cluttered screen and improve the computation speed of the system when handling large number of factors in the situation model.

7.5.1.3 Resulting higher comprehensive insight

The analysis result indicates that the participants were able to gain a significantly higher extent of comprehensive insight in the treatment session compared to the control session. This result implies that the proposed system is capable of enhancing users' ability to conceptualize and comprehend the overall picture of analytics problem. Such holistic understanding of the problem situation can be achieved by comprehending the interrelationships between the factors that are meaningful at the problem-solving

level. As an example, in the user study setting, the comprehensive insight refers to the holistic understanding on the stock market as an ecology which is made up of the relationships between the stock prices, companies' financial information, industrial trends, and macroeconomic factors. The "assisted situation modelling" design principle theorises that the users can be more effective in terms of conceptualizing and comprehending the holistic view of the problem when they are supported by the comprehensive features to develop external representations of the situation model.

Follow-up testing indicates that there is a medium-strength positive association between the average number of unique factors in the participants' situation models and the comprehensive insight, $r = 0.437$, $p = 0.016$. However, the number of situation models created has no association with comprehensive insight. These findings suggest that to achieve the holistic picture of the problem situation, the capability to include wider variety of factors in the situation model (such as industrial factors, economic factors, and company specific financial factors) is more important than the ability to create multiple parallel "possible" models of the problem situation. Such findings also suggest that not all comprehensive features proposed equally contribute to the comprehensive insight. Moreover, tests against demographics confirmed that the enhanced comprehensive insight observed among the participants are not a function of their demographics.

With the overall evidence, this study concludes that the design principle "assisted situation modelling" has a positive impact on the users' comprehensive insights. It enhances the user capability to understand the relationships between the key factors in the problem situation and to conceptualize the structure of the problem situation. The result is in line with existing studies that show that external representation of the problem model allows the users to offload their working memory to focus on the actual analytical activities such as conceptualization, internalization, and reasoning (Ayres & Van Gog, 2009; Kirsh, 2010). At the time the experts alter each parameter in the situation model in the proposed system, the model is capable of providing instant feedback to reflect how change on the big picture of the problem. Such rich two-way interaction allows the users to continuously fine-tuning the model to reflect their train of thought, thus enabling to them to gain deeper understanding of the dynamic nature of the problem. As a result, such interaction improves their understanding of the problem situation.

7.5.1.4 Resulting higher predictive insight

The analysis result indicates that the participants were able to gain significantly higher extents of predictive insight in the treatment session compared to the control session. This result implies that the proposed system is capable of enhancing the users' ability to project the future behaviours of problem situation. This study theorised that the predictive insight can be effectively achieved when the users are supported 1) with a default predictive model which uses historical data to predict the future states of key factors in the problem situation based on their interrelationships, 2) then with the capability for the users to override the predictive parameters (i.e. the likelihood of different states for each of the key

factor), and 3) with a semi-automated forecasting algorithm when the users are uncertain about the future states of the key factors. These features derived from the design principle "enabling predictive reasoning" are theorised to being capable of supporting the predict & simulate activity in the knowledge actualization phase of data analytics.

However, the present data does not reveal any significant association between these features and the predictive insight. This study believes that this is because the users' predictive insight resulting from overall interaction between these features is larger than the sum of the individual interactions with each of the features. In this sense, the effects of such interaction may not be able to be captured by interactions with individual features. On top of that, the improved predictive insight among the participants might be the effects of other features which are not theorized to have an effect on predictive insight. In accordance with this, this study found a positive association between predictive insight and with a supporting feature for the users to evaluate the accuracy of the predictive model, $r = 0.401$, $p = 0.028$. This may imply that to enhance predictive insight, the ability to gauge the accuracy (i.e. fitness) of the predictive model possibly is more important than the ability to fine-tune the default predictive model.

Overall, the design principle "enabling predictive reasoning" is found to enhance the predictive user capability to predict the future behaviors of key factors in the situation model. Nevertheless, caution should be taken as there is no evidence to show the effects have directly resulted from the prediction features. The observed enhancement in the predictive insight could be the effects of overall interaction that is unobservable in the separate feature. This overall effect is reflected in the significantly greater predictive insight gained in the treatment session in comparison to the control session. This study believes that the value of the predictive reasoning features is that the prediction and simulation are shaped by the unique way that each user conceptualizes their problem situation. Unlike pure number-crunching predictive techniques, predictive reasoning in this study involves motivated, continuous, and interactive effort to anticipate the trajectories of the key factors in the situation model. In other words, it allows the users to perform a variety of "what if" analyses to test their speculation of what is going to happen and how it is going to influence the overall problem landscape. Therefore, the design effectively supports the prediction & stimulate activity and enhances the predictive insight gained by the users.

7.5.1.5 Resulting higher prescriptive insight, not without unintended effects

The analysis result indicates that the participants gained significantly higher prescriptive insight in the treatment session than in the control session. This implies that the proposed system is capable of enhancing users' ability to assess the impacts of their potential course of action on the future states of the problem situation in the light of their objectives and constraints. It is theorized that such insight is made available by allowing the users 1) to optimize the resource allocation which can meet the conflicting objectives with their constraints while compensating for the risks attributed to uncertainty and 2) to accurately and rigorously assess the risks associated with the course of action. In this regard,

features which are derived from the design principle "enabling stochastic optimization" were built into the proposed system to support prescriptive insight.

The features are found to be significantly associated with prescriptive insight, $r = 0.571$, $p = 0.034$. The prescriptive insight is not caused by the variances among the participants. In accordance to the evidence, this study concludes that the design principle "enabling stochastic" allows the users to be more effectively in evaluating the impacts of their potential course of action in relation to the objectives and constraints. As a result, the users would able to gain higher extents of prescriptive insight. Prescriptive insight is the highest state of knowledge about the problem situation that suffices to allow the users to make decisions. With the optimized course of action and the understanding of its risk, the users would be able to make informed decision that will meet the conflicting objectives and satisfy the constraints, while minimizing the risk caused by the uncertainty. The prescriptive insight that is built on top on the previous insights allows the users to make an informed decision to solve the analytics problem.

As an additional finding, the observation in this study reveals that the participants were not entirely rational. More than a quarter of the participants allocated their capital (nearly) evenly to the 4-5 selected stocks, rather than proportionate to the suggested portfolio allocation. This study has speculated that the users may not adopt the suggested portfolio allocation as it is, and might adjust the allocation. However, this study did not expect the users would allocate their capital evenly across the selected stocks, after reviewing the suggested portfolio allocation. Follow up interviews with the participants revealed that the participants believe that equally allocating the capital across the stocks will reduce the change of losing large amount of investment if their prediction were wrong, and thus presumably can reduce the risks of loss. This behavior is showing that the participants attempted to follow their intuition and neglected the optimized suggestion. Such allocation defeats the purpose of the design principle "enabling stochastic simulation" to help the users overcome biases. However, the control over whether the users should adopt the optimized suggestion is beyond the system design. It is believed that training and practice may increase the users' awareness of the importance of the optimized suggestions.

Moreover, several participants in this study have commented that they felt "*being fooled* (misled)" when their investments resulted in a loss as a result of following the suggested portfolio allocation. Note that the moderator has explicit told every participant that the accuracy of the suggestion is only as good as the historical data and the model they built. Based on the participants' reactions, this study conjectures that suggestive recommendations given by the system may cause the participants to take the figure (i.e. in terms of number of stock to invest) with overinflated confidence. As for the improvement, the optimization results can be presented in the form of relative strength (i.e. graphical bars) rather than exact number of stocks / amount of capital.

7.5.1.6 Resulting lower perceptive insight

The analysis result indicates that the participants have gained lower perceptive insight in the treatment session in comparison to the control session. This result implies that the alternative system is more effective in supporting users to perceive and understand the relevant information elements. This study theorised that the perceptive insight can be enhanced by 1) supporting the users to collect and manage the observations made from derived from information elements and 2) supporting the visualization and analysis of the meta-observation information. These are the two features derived from the design principles "enabling managed observation" and "enabling exploration convergence". In order to have a fair comparison between the proposed system and the conventional data analytics system, the alternative system is a full-blown visual analytics system. As a full-blown visual analytics system, the alternative system has the capability to let the users build custom visualization from scratch. On the other hand, the feature is stripped from the proposed system.

This study originally conjectured that the effects of the design principle "enabling managed observations" could offset the lack of such feature. Nevertheless, the results have proven this study's conjecture is wrong. The capability for the users to build custom interactive visualization is the core of the data exploration. The removal of the interactive visual explorer significantly impairs the ability of the users in understanding the information elements, and cannot be offset by the features from the design principles "enabling managed observations" and "enabling exploration convergence".

Despite this, this study tried to find out whether the features can have positive effects on the participants within the treatment session itself. By independently analyzing the perceptive insight in the treatment session, there is no correlation between the perceptive insight and the features derived from the design principles. Based on these findings, this study concludes that the two design principles do not have the theorized effects on the participants' perceptive insight. This study believes that the features did not directly enhance the way the users derive observations from the data, but they provide necessary components for the high-level analytical activities, such as developing the situation model of the problem situation. The features from both the design principles enable the users to easily and accurately retrieve the observations to be used in the information synthesis and knowledge actualization phases.

7.5.2 Findings pertaining to the value of the analysis outcome

Key findings: At the overall level, the participants perceived the value of the analysis outcome from the treatment session as significantly higher than the analysis outcomes from the control session. Among the 12 dimensions, the dimensions which were perceived to be significantly higher compared to the control session are 1) uniqueness, 2) robustness, 3) knowledge building, and 4) applicability to

decision making. On the other hand, the analysis outcome from the control session were perceived to be more understandable than from the treatment session.

7.5.2.1 Resulting analysis outcomes with higher overall value

The analysis outcomes from the treatment session were generally perceived to be of higher value than those from control session. This finding suggests that the proposed system would be able to assist users in deriving an analysis outcome with higher value. The higher the value, the more desirable is the analysis outcome in the users' problem-solving domain. And thus, the more likely that the analysis outcome would be actually deployed into the physical world. In contrast, analysis outcomes with low value might never go beyond the analysis stage and never get presented to the decision boards. As such, this study conjectures that the proposed system would be able to increase the throughput rate of the analytics system. The system increases the success rate of turning analyses into insights that are valuable to the domain. The following paragraphs provide more details into which dimensions have played a more important role in contributing to the overall value enhancement.

7.5.2.2 Resulting more unique analysis.

The participants have perceived the analysis outcomes from the treatment session to be more unique. In other words, they believe that the analysis outcomes cannot be easily imitated by others. This result suggests that the proposed system allows users to derive uniquely important analysis outcomes. Unique analysis outcomes can be achieved because the proposed system allows the users to incorporate their domain knowledge to develop a custom-made predictive model. The predictive models collate, organize, formalize, and finally combine the participants' domain knowledge together with the historical data to produce domain-meaningful models. The predictive models in turn are used to customize the simulation and risk optimization algorithms in the later phase of the analysis. As a result, the analysis outcomes from the proposed system have factored in the differences in the individual users' accumulated knowledge, judgement, and personal experience.

It is commonly agreed that uniqueness is a highly appreciated value of analysis outcomes, particularly in business-related domains (Pavel & Dragos, 2010). Unique analysis outcomes are desirable because they could be translated into strategy which can be leveraged for sustainable advantage. It allows the decision makers to ride the edge of a new trend and to benefit from the pioneering advantage. Such a head win is important to stock market and other business strategies (Jarzabkowski & Wilson, 2006). In the stock market, it allows the investor to take advantage of a new trend to maximize their earnings, before the majority floods in to diminish momentum of the price trend. With the capability of the proposed system, the individual or organizational users who adopt the proposed system can be less worried about losing the differentiation advantage even if their competitors

adopted the same data analytics system. The uniquely created models become their intellectual assets, which is difficult to imitate, in comparison to the other resources such as datasets and analytics skills.

7.5.2.3 Resulting more robust analysis.

This study found that the participants have perceived the analysis outcomes from the treatment session to be more robust than the control session. This implies that the participants believe that the analysis outcomes are not overly susceptible to underlying data change and assumptions being void. The proposed system allows the users to derive robust outcomes by enabling them to rigorously assess their assumptions, uncertainty, and situation model. The users can create multiple versions of future scenarios. Each of the versions can be used to represent different possible combinations of assumptions and uncertainty. Subsequently, these future scenarios are used to assess the effects of different potential courses of action. Moreover, the stock allocation solution suggested by the proposed system is optimized against the uncertainties in each future scenario to maximize the earnings while minimizing the loss. As a result, the users would have a better grasp of how their decision would perform under different circumstances, and hence, have higher confidence in their analysis outcome.

Analysis outcomes with low robustness might be a sign of stringent or unrealistic assumptions. This may cause the analysis outcome to be very appealing on the paper, but the value will be greatly diminished the moment it is brought into a practical setting. Therefore, robustness may also be an indicator for the practicality of the analysis outcomes. In contrast, robust analysis outcomes are highly appreciated by decision makers, especially in fast changing and highly uncertain domain. The higher the robustness, the more confidence the analysts have in the outcomes. Study has stated that a robust analysis outcome enhances the potential for the decision makers taking change action (Thomas et al., 1993). The visual analytics community has set dealing with uncertainty and robustness in the analysis outcomes as one item on the key research agenda. This study believes that it has contributed a small step toward this objective. The proposed system allows users to interactively deal with the uncertainty in the model and the data, hence, allowing them to produce analysis outcomes that are robust.

7.5.2.4 Resulting more knowledge building

In the user study, the participants have perceived that the knowledge building value of the analysis outcome from the treatment session is higher than from the control session. This finding implies the proposed system can better help users to gain new knowledge about their problem situation, compared to conventional data analytics system. Essential to the knowledge building is the construction of schemata, that is, a knowledge structure organized around a central concept. In the construction process, working memory plays a major role in reasoning, organizing, and integrating the knowledge before the knowledge is consolidated in the long-term memory. Working memory can be a major impediment to knowledge building; this is especially true when dealing with a complex analytic task which contains a

high number of new and interconnected information elements. The proposed system scaffolds this construction process by allowing the users to build an external representation of the schemas, that is, the situation model. Such external representation enables users to overcome the capacity limitation of working memory. The externally presented schemas offload the users' working memory to better focus on the analytical reasoning and internalization, and hence, result in better knowledge building (Yedendra B. Shrinivasan, 2010). Moreover, the situation model is capable of providing instant feedback when the users alter each parameter. As a result, the users are able to engage in a rich-form two-ways interaction that greatly facilitates knowledge building.

Knowledge building has been widely agreed to be a crucial prerequisite for solving complex problems. The continuously evolving nature of problem solving requires the analysts to progressively learn the nature of the problem before they able to develop the solutions (Kirsh, 2009; Mirel, 2004). Knowledge building enables the users to establish connections between subtle information and to apprehend the inner nature of the problem. Moreover, knowledge building entails a mental shift in a person's perception of a problem situation. This study believes that the proposed system not only improves knowledge building for the current problem solving, it also enriches the users' mental model during the analytics process. At a consequence, the users' capability to solve similar class of the problem in the future could be enhanced.

7.5.2.5 Resulting in more applicable outcomes

The result indicates that the participants found that the analysis outcomes from the treatment session can be applied more directly to support their decision making. In other words, the proposed system produces analysis outcomes that are closer to the problem-solving level. By supporting the higher-level analysis phases, namely information synthesis and knowledge actualization, the mental mapping between the analysis outcomes and the decision is clearer. This it allows the users to feed to the outcomes as inputs to the decision-making process. On the other hand, given that the alternative system does not support the high-level analytical activities in data analytics, the produced analysis outcomes are often low-level technical information. The mental gap between the analysis outcomes and the problem solving is therefore wider. A common challenge for the users is to further process the low-level information in order to determine how they can apply the technical information to support their decision.

Observations during the user study shows that, in attempts to overcome the mental gap, the participants in the control session tended to carry out the analytic processes outside the system, either entirely in their mind or with the aids of paper and pen. The participants commonly try to gain a big picture from the individual findings by jotting down a tentative list of stocks. Then the participants constantly add, edit, or delete the stocks in the list as they discovered new information. A common

activity observed among the participants is that they rate each stock in the list with a single rating. Different participants use different symbols to represent the desirability of each stock. Some used “star” shape, while the others use “tick” and “circle” shape. The greater number of these symbols indicates the more desirable is the stock. This observation shows the participants attempted to use a single qualitative rating to represent their judgement on multiple criteria in an ambiguous and potentially biased fashion. All criteria are treated by the participants equally, although in reality some of these criteria supposedly weighted more than the others. No participants have used techniques that are similar to a “weighted criteria decision matrix” for a more rationale rating.

At the end of the analysis, based on the single rating of each stock, the participants decided the proportion of the investment capital that goes into the top 4 or 5 most desirable stocks. Yet most of the time, the capital allocation did not reflect with ratings that the users have given to the stocks. These behaviors suggest that the participants tended to rely on intuition and to leap to the conclusion without a systematic approach. When asked, the participants often failed to clearly justify their actions. Or if they were able to explain their actions, the justifications seem more like post-factual reasoning that tried to rationalize their action. Therefore, this study believes that the proposed system is not only capable of enhancing the mapping between the analysis outcomes and the problem solving, it also induces rationality on the users. It enables the users to carry out structural analyses which are more systematic, traceable, and less susceptive to cognitive bias.

7.5.2.6 Lower Understandability

The result indicates that the participants found that the analysis outcome from the treatment system is significantly harder to understand and interpret. The lower level of understandability in the treatment session is probably was caused by the comprehensiveness of the analysis techniques used in the analysis. The techniques used in the treatment session requires the participants to interpret probability distribution chart, cross-tabulation tables, and other statistical indicators which are not common for general users. This is also supported by the observations that during the practice session, one of the most common questions from the participants was about how to interpret the results displayed in the module 2 and 3. On the other hand, the analysis outcome in the control session is presented in commonly used visualizations such as pie chart, line graph, histogram, and scatter plot. Hence, it is not surprising that the outcome is easier to be interpreted.

The understandability defines the degree to which the participants were able to comprehend the analysis outcomes in the context of their problem. Outcomes that are complex and overwhelming are difficult for the users to interpret in the context of their problem, such as the task objective, constraints, and rules. This in turn prevents the users from effectively understanding and solving the problem situation. This study believes that this is one of the key improvements which should take the very first

priority. The comprehensiveness of the analysis outcomes can be customized according to the users. For instance, the analysis outcomes can be simplified and summarized as a single indicator. The details of the analysis outcomes should be made optional, only being displayed on request. Such a configuration will allow the novice users to intuitively understand the outcomes, while allowing the expert users to have access to the detailed and technical outcomes.

7.5.3 Findings pertaining to the decision performance

Key findings: Mixed results were found in term of decision performance. The decision performance was assessed from both quantitative and qualitative aspects of the participants' decision. In terms of the objective measure, the total earnings from the treatment and the control session are not statistically different. In addition to total earnings, this study also examined the overall change in the participants across the two sessions. The number of participants who have positive total earnings in the treatment session is 17% higher than the control session. Nonetheless, when looking at the earnings above random baseline, the number of participants who earned above the random baseline in the treatment session is 10% lower than the control session. In term of the qualitative aspect of the participants' decision, the results suggest the participants' stock selections in the treatment session have significantly higher degree of match with the stock selections of domain experts.

Overall, the evidence in this study indicates that the proposed system used in the treatment sessions has no significant improvement on the quantitative aspect of the users' decision. However, the qualitative aspect of the users' decision has been improved. As such, this study concludes that, as opposed to the proposition, the proposed system did not definitively improve user's actual decision performance compared to a conventional system. Nevertheless, the proposed system enables average users to derive analysis outcomes that are relatively closer to experts' decision, in comparison to conventional analytics systems that do not support information synthesis and predictive simulation.

7.5.4 Findings pertaining to Usability

Key Findings: Analyses pertaining to the dimensions of usability have indicated that the participants found the proposed system was significantly harder to use and more difficult to learn, compared with the alternative system. Despite that, such challenges did not significantly reduce the user satisfaction of the participants in using the proposed system. The participants perceived high extents of user satisfaction (with average scores of above 4 out of 5) in both the treatment and the control sessions. More importantly, the participants generally found the proposed system is more useful than the

alternative system. There is no difference between the participants' intentions to continuously use the proposed system and the alternative system.

7.5.4.1 Harder to Use and Learn.

It is expected and reasonable that the participants found the proposed system was harder to use and more difficult to learn in relation to the alternative system. The proposed system is more complicated given that it 1) contains greater number of features, 2) requires specific interactions, and 3) requires basic understanding of the analytic concepts. The proposed system contains two more modules in addition to the data exploration module in the alternative system. Each of these modules contains multiple analytical functions which can be executed in non-sequential and repetitive order. Moreover, each function requires specific interactions which may not be familiar to general users. Observations recorded that 76% of the participants from the treatment sessions have asked for assistance on how to execute certain functions at least once.

The biggest hurdle for the participants was to learn the fundamentals about the analytic concepts used in the modules that are exclusive to the proposed system. This includes the "why" and "how" to use certain functions and how to interpret the outputs. While the alternative system uses common analyses such as percentage and ratios, functions in the proposed system involve advanced analytic concepts which are not common for general users. For example, a statistical concept such as probability distribution is new and not easy to be comprehended by untrained users. As suggested by studies in cognitive learning, information that is entirely new to the learners imposes heavy cognitive loads on them, and hence, reduces their capacity to learn. With consistent evidence pointing to the fact that the proposed system is hard to use and learn, one of the most important improvement for the system is to improve its user friendliness.

7.5.4.2 User training is key to better utilization, also a revenue model

Regardless of the proposed system being perceived as harder to learn and to use, 83.3% of the participants were able to complete the analysis by using all three modules. The remaining 17.7% of the participants did not use the third module – "predictive simulation". While the most common reason given is "*not enough time*", the second and third common reason are "*don't think it's necessary*" and "*I'm confident with analysis*". A follow-up analysis has found that the participants who perceived the proposed system to be easier to learn have spent significantly more time on modules 2 and 3 of the system, M = 11.9%, 95% CI [1.733, 22.10], p = 0.23, d = 0.88. This implies the importance of the user training. Once the users undergo proper training, very likely they will have better acceptance of modules 2 and 3 in the proposed system. The importance of user training could also suggest that user training can be one important part of the business model, where tailored trainings and webinars can be delivered

to generate constant stream of revenue besides the software sales. Similar business model has also been seen in NVivo, IBM SPSS, and SAS Enterprise Miner.

Besides examining the usability of the proposed system from the perspective of overall participants, this study also seeks to understand how the participants' demographics could influence how they perceive the usability of the proposed system differently. Along this line of examination, a series of tests revealed that domain knowledge and domain experience of the participants had effects on how perceive the usability of the proposed system. The following paragraphs discuss the findings in detail.

7.5.4.3 Not Dummy-proof, Domain knowledge is important

A one-way ANOVA test has discovered that the participants with higher level of financial knowledge perceived the proposed system to be more useful than participants with lower level of financial knowledge perceived, $p = 0.029$. However, the level of investment knowledge did not influence how the participants perceived the usefulness of the proposed system. Consider investment knowledge as the "hard domain knowledge" and financial knowledge as the "soft domain knowledge". Investment knowledge is concerned with very specific and technical knowledge in carrying out stock market analysis. In contrast, financial knowledge is a broader knowledge aligned towards subjective understanding of the financial concepts, including understanding balance sheet, gauging financial health of business, and understanding effects of compound interests.

The proposed system largely focuses on high-level analysis to understand the effects of high-level factors such as bank interest rate, foreign investments, and industrial-specific indexes on the stock price trend. Given this, the participants who have a higher level of financial knowledge can better take advantage of their knowledge to include sensible factors into their situation model, to build domain-meaningful models by connecting the factors in a realistic way, and to be more proficient at interpreting the outputs provided by the proposed system into the context of their analysis. Other studies have also stated the importance of domain knowledge in data analytics to obtain analysis outcome that is meaningful in the domain (Heer & Shneiderman, 2012a). In contrast, technical-oriented knowledge may not as useful for taking the advantage of the system. Similarly, these are similar to the circumstance where the broad domain knowledge of crime investigators would allow them to take advantage of an intelligence investigative system better than his or her counterparts who is less expert in the domain. This finding is aligned with the goal of this study that is not to make the proposed system a dummy-proof tool, but as a tool that can augment the human experts' analytic processes.

7.5.4.4 Low resistance from inexperienced users

Two independent groups are formed by dividing participants who have investment experience (46.7%) and who do not (53.3%). Two groups of participants have rated the usefulness of the proposed system as very high (above 4.2 out scale of 5). Nevertheless, an independent-sample T-test has revealed that participants who have no stock investment experience perceived that the proposed system is more useful than their counterparts who have investment experience. The difference was statistically significant, M = .402, 95% CI [0.046, 0.757], p = 0.023, d= 0.474. This finding may imply that while the proposed system did not receive significant resistance from the experienced users, it was also being well-accepted by users who have no investment experience.

7.5.4.5 Higher domain knowledge, Harder to use?

Interestingly enough, this study found that participants with a higher level of financial knowledge perceived the proposed system to be significantly harder to use, compared to their counterparts who have a lower level of financial knowledge. The observations recorded show that participants with a higher level of financial knowledge used power-user features (see appendix for the list of power-user features), while their counterparts tended to stay with default configuration and often skipped power-user features. Power-user features often require more complex inputs, involves highly customizable parameters, and contain more technical jargon in the outputs. In that regard, the user experience in using the power-user features is believed to have contributed to the lower ease of use. As the implication of this finding, more design efforts should be allocated to improve the usability of these power-user features. The power-user feature should be redesigned to provide a positive user experience for the existing users, while encouraging users with lower level of domain knowledge to use the features.

7.5.4.6 Summary of Discussions on the System's Usability

Overall, given the complexity of the system, user training plays an important role in encouraging the users to take better advantage of the advanced features in the system. Meanwhile, the user interactions of the system needed to be simplified to improve its user friendliness. This is particularly true for the power-user features. On the positive side, users with or without actual investment experience were highly open to the proposed system, especially the users who are fresh to investment.

7.5.5 Finding pertaining to Cognitive Load

Key findings: two out three types of cognitive load were found to be significantly higher in the treatment session compared to the control session.

7.5.5.1 The task in the treatment session is more taxing, by its nature

The participants indicated that they have experienced a higher intrinsic cognitive load in the treatment session. Intrinsic load, associated with contents of the tasks such as working memory needed to store

and process the information elements, is influenced by both the number of information elements and the interactivity of those element. This study did not anticipate this result because both the treatment and the control sessions use the same datasets. However, the results imply that even with the same datasets, the participants might have perceived that the contents in the treatment session were more taxing than those in the control session. One possibility of such an occurrence is that the proposed system exposed the interactivity of the information elements which otherwise oblivious. With the awareness of the complex interactivity, the participants felt that the datasets that they received in the treatment session was more complicated. On the positive side, it may indicate that the proposed system can help users to discover hidden interactivity between information elements. On the other hand, the participants might have overwhelmed by the possible connectivity between the information elements. The present proposed system first relies on the user's domain knowledge to connect the semantically related information elements. Then, historical data is used to evaluate the validity and strength of the connections. However, this human-oriented method may be ineffective when there is a massive number of interactivity, which could also be imposing a higher load on the users' cognition.

7.5.5.2 The system design induces unnecessary load on the participants' cognition

A significantly higher extraneous cognitive load in the treatment session might suggest that the designs of the proposed system can still be largely improved. Extraneous cognitive load refers to the load induced by the design of the system, in terms of how the information is presented and how the user interactions is designed. In that sense, extraneous load is caused by activity or information that does not foster knowledge building. For example, efforts spent on seeking for important among information clusters that contain largely irrelevant information elements will contribute to extraneous effort. Previous studies found that having a surplus of information presented to the users at once is one of the main factors of the extraneous load. As such, this finding may suggest the proposed system is conveying excessive information elements to users. This is true in the sense that the proposed system displays all key outputs and less-essential outputs at once. This design can be found in the result dialogs that shows the fitness of the predictive model, the optimized allocation of investment based on simulation, and the converged observation dialogs. Besides the presentation of the information, the higher extraneous load in the treatment session could also be the result of usability of the user interface. The harder-to-use interface costs the participants higher cognitive efforts that are not going into the knowledge building, but into activities that do not contribute to the analytic goal. More detail about the user interface usability issue have been discussed in previous subsection 7.5.4.

7.5.5.3 More effective learning?

The higher germane load perceived by the participants in the treatment session is an indication that greater effective efforts have contributed to the knowledge building in comparison to the control session.

Germane load is associated with the mental efforts that are directly relevant to knowledge building activities, such as constructing and refining mental schemata from the information elements (Vandewaetere & Clarebout, 2013). It is commonly theorized that the higher extraneous load will reduce the pool of cognitive resource from being used for knowledge building. Based on this assertion, the control session which has a significantly lower extraneous load should have a higher germane load. Nevertheless, this study found that the germane load in the treatment session remained higher than in the control session. This study conjectures that the supposedly low germane load in the treatment session has been offset by the features to support knowledge building. Features such as situation modeling have potentially offloaded the cognitive resources which required to hold the mental schemata, thus making those resources available for knowledge building. From the other perspective, this also implies that the full potential of the proposed system is not achieved. If the considerably high extraneous load can be reduced, even more cognitive resources will be available to be used for knowledge learning. This is another indication pointing to the importance of improving the usability of the proposed system.

7.5.6 Relationships from Insights to Decision Performance

Theoretically, a higher level of overall understanding about the problem would result in analysis outcomes with higher value. However, the follow-up regression test did not discover a significant relationship between the overall insight and the perceived quality of analysis outcomes. Interestingly, there is a significant positive relationship between high-level insights (insight level 2 and 3) and the perceived quality of analysis outcomes ($p = 0.004$, Beta = .446, Adjusted $R^2 = 23\%$). This finding may suggest that the more the participants understand the problem as a whole and predict the effects of potential course of action on the future scenarios, the more likely it is that the participants will able to produce analysis outcomes which can delivery higher value to the domain. This finding provides support to this study's proposition that most of the existing business analytic systems often fail to produce valuable insight because the high-level cognitive processes are not explicitly supported.

Similarly, it is theorized that a higher value of analysis outcome would lead to higher decision performance (Endsley et al., 2011). The regression result in this study shows there is no clear relationship between the value of analysis outcome and either the quantitative or the qualitative aspects of the decision performance, $p = 0.159$ and $p = 0.231$ respectively. It is noteworthy that decision performance is a complex function which is often influenced by other factors that are beyond the control of the study design. In practice, one's decision performance could be affected by pressure, boundary of decision making authority, resources, social influences, and even a sense of randomness. Even if the perfect information is available, it does not guarantee a perfect decision performance. As in this study, there is no perfect relationship between the information and the actual price trend. The actual price trend can be the result of complex market transactions, irrationalities of investors, political turmoil, randomness, and other factors which are not unsuspected to influence the price.

7.6 Implications of the results on the Propositions

Recall that the design framework introduced in Chapter 7 contains the propositions theorizing that the design principles would be able to increase the user performance in corresponding problem-solving activities. *Figure 86* shows whether the results of analysis are able to support the propositions.

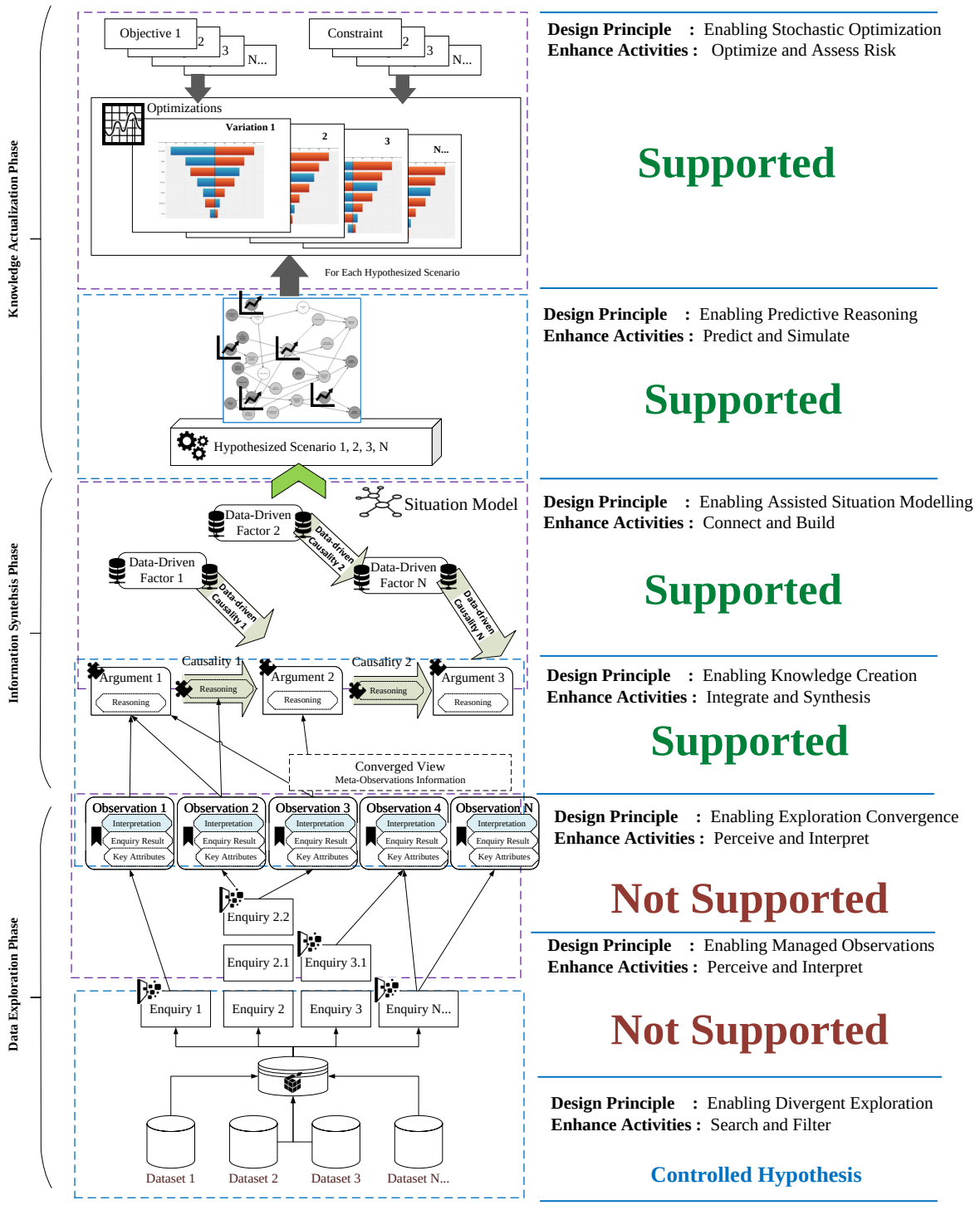


Figure 86. Propositions and Results

The findings suggest that the data analytics system that incorporating the design principles is more effective at supporting the problem-solving activities during the information synthesis and knowledge actualization phases of data analytics, when compared to conventional data analytics systems. Specifically, the proposed data analytic system can enhance the user performance in 1) integrate & synthesize and 2) connect & build activities during the information synthesis. The proposed system is also capable of enhancing the user performance in both predict & simulate and optimize & assess risk during the knowledge actualization. In other words, the designs can help to reduce the gap between low-level technical analytic results and the high-level conceptual understandings which are required to solve the analytics problem. Therefore, the users are more likely to use the insights gained from the data analytics to inform decision and actions.

However, the findings suggest that the data analytics system that incorporated the design principles is not significantly more effective at supporting the problem-solving activities during the data exploration phase. Specifically, the design of the proposed system does not enhance the user performance in the perceive & interpret activity, when compared with the conventional data analytics systems. Note that, the effects of the design on the "search and filter" activity were tested as a controlled hypothesis to ensure the effects are due to these designs rather than to other confounding factors. Moreover, it is required by the user study setting that the design for supporting search & filter should be implemented in both the proposed system and the conventional data analytics to enable the participants to have the same fair starting point in the data analysis.

Chapter 8
Conclusion

8.1 Overview

This chapter first provides a summary of this study in Section 8.1. Then the theoretical and practical contributions of this study are presented in Section 8.2. The limitations of this study are discussed along with some suggestions for future research in Section 8.3. Lastly, a conclusion is presented in Section 8.4.

8.2 Research Investigation

During complex data analysis tasks, data analysts often engage in a series of problem-solving activities, including extracting semantically meaningful information from the data, synthesizing information to form higher-level concepts, creating a mental depiction of the problem, and mentally simulating the impacts of possible scenarios (David & Michelle, 2009; J. Kohlhammer et al., 2011).

Nevertheless, most of the existing data analytics systems to date are the result of advancement in data-driven and computational techniques. These systems were designed with little consideration of how users behave in a complex analytical task. As a result, these systems often focus exclusively on low-level data exploration and fail to effectively support thee high-level problem-solving activities. Consequently, analysts have to contend with the gap between the low-level technical analytic results and the high-level conceptual understandings which are required to solve the analytics problems. Thus, data analysts often find it challenging to determine how the analytic results can be used to inform their decision-making and solve their real-world problems. In other words, existing data analytics systems often fail to deliver the actionable insight promised.

The goal of this study has been to develop a design of data analytics systems that can explicitly support the analysts' problem-solving activities. This study hypothesizes that when the problem-solving activities are supported, analysts are more likely to produce higher-level insights that are more readily able to inform decision-making. In order to inform the design, this study asserts that there is a need to first understand and define actionable insight. This is because the existing understandings of the term "actionable insight" are generally too vague and abstract. Without exactly knowing what to achieve, the design would hardly be useful. In order to achieve this design goal, this study asserts the need for 1) systematically understanding and defining actionable insight, 2) understanding the problem-solving activities and outcomes required to achieve actionable insight, and 3) proposing system features that can effectively support the problem-solving activities.

This study employs design science research as the methodology to guide its overall design. Through an integrated understanding of relevant theories, namely situation awareness (SA), sensemaking, and complex problem solving, this study conceptualizes actionable insight as a multi-component construct that is comprised of three major insights: analytic insight, synergic insight, and prognostic insight. Such conceptualization enables this study to understand and define actionable insight. Additionally, a conceptual explanatory framework is developed based on the conceptualization of actionable insight. As a whole, the framework provides a holistic explanation of the complex analytical tasks. It explains the information processes, user behaviors, cognitive states, and reasoning outcomes in the different phases of data analytics. In other words, it enables this study to understand the processes and components required for data analysts to achieve actionable insight. More importantly, the framework provides systematic and theoretically-grounded design requirements for improving the design of the data analytics systems.

Based on the design requirements, a conceptual design framework was developed. The design framework comprises a set of design principles. The objectives of the design principles are 1) to provide high-level conceptual design to address the design requirements and 2) to acts as the blueprint for translating the conceptual design into tangible system features. This study conjectures that, together, these design principles can be used to inform the design and development of data analytics systems that can effectively support the problem-solving activities in complex analytical tasks.

To evaluate the effectiveness of the proposed design, a prototype system was developed based on the design framework. The prototype system was evaluated against a conventional data analytics system in the user study. A user study involving 30 participants was undertaken in a controlled setting. Each participant was required to participate in two sessions, one using the prototype system and one using the conventional data analytics system. In each session, the participants were required to analyze a stock market and to develop profitable stock portfolios. This was followed by a self-administered questionnaires survey, and a follow-up interview.

The analysis of the data collected from the user study shows that the proposed design has effectively supported the participants' problem-solving activities in the information synthesis and knowledge actualization phases, but not in the data exploration phase. Additionally, the proposed design was found to increase the perceived quality of the analytical result, implying that the result is more likely to be deployed into the physical world through decision making. Lastly, mixed results were found in the analysis of the effects of the proposed design on actual decision performance. The qualitative aspect of the participants' decisions has been improved, but the quantitative aspect of the decisions was not improved. In terms of system usability, participants generally found that the prototype system is more useful for supporting them to carry out their task, but it is also significantly harder to use and to learn. Overall, the findings suggest that the proposed design can support users to achieve higher quality and

more complete insights. In other words, the design enables the users to gain better understanding about their problem situation, and thus reduces the gaps between the low-level technical analysis and the high-level understanding required for solving analytical tasks. However, the design does not promise an improvement in the actual outcome of the decisions.

To conclude this section, the research questions from chapter one are restated and answered based on the findings.

1) How can actionable insight be systematically defined?

2) What are the processes and requirements to achieve actionable insight?

3) How can these processes and requirements be effectively supported?

8.2.1 Answering Research Question: How can actionable insight be systematically defined?

This study proposes a definition of actionable insight based on the conceptual explanatory framework. The framework conceptualizes actionable insight as a multi-component concept which consisting of three major components, namely analytic insight, synergic insight, and prognostic insight.

Table 37. Components of actionable insight

Major Insights	Description
Analytic insight	Understanding and interpretation of individual analytical results
Synergic insight	Comprehension of the connections between the analytic insights and understanding of the problem situation as a whole
Prognostic insight	Prediction of the problem situation's future states and the assessment of their effects on the problem situation, objectives, and constraints

Based the conceptualization, this study defines actionable insight as:

> *A set of progressive knowledge about the analytics problem, based on prognostic insights derived from synergic understanding of analytical results, which enables the user to make an informed decision to solve the analytics problem.*

Based on this definition, actionable insight is the coherent states of knowledge the users gained at different data analytics phases. Together, these progressive understandings provide the users with sufficient understanding of the problem situation in order to decide on a solution that best suits the user's objectives and anticipated scenarios. This thereby constitutes the "actionable" notion of the term.

8.2.2 Answering Research Question: What are the processes and requirements to achieve actionable insight?

Besides being used for conceptualizing actionable insight, the conceptual explanatory framework itself provides a holistic understanding of complex analytical tasks. It explains the information processes, user behaviors, cognitive states, and reasoning outcomes in the different phases of data analytics. The framework allows this study to identify and understand the problem-solving activities and components required to achieve actionable insight.

Based on the way "actionable insight" is being conceptualized, components required to achieve actionable insight are the specific insights that form the actionable insight. To facilitate greater explanation, the three major components of actionable insight are further decomposed into six specific insights, namely identification insight, perceptive insight, integrative insight, comprehensive insight, predictive insight, and prescriptive insight. These insights are derivatives from the constructs the situation awareness (SA) theory. The theory refers to the different states of situation awareness as the cognitive outcomes of problem-solving activities that are directed toward actionable insight. This study uses derivatives to specifically describe the information-processing states that result from the human-information discourse in data analytics.

Specific problem-solving activities are required to achieve each of the insights. The problem-solving activities root their theoretical basis in sensemaking theory, which was developed to understand how humans process information for complex problem solving. It focuses on the information-processing activities, particularly in the field of intelligence analytics. This study adapts activities for explaining the problem-solving activities in data analytics, without altering its theoretical doctrine. *Table 38* shows how these problem-solving activities and their target insights correspond to the three main phases of data analytics.

Table 38. Components of actionable insight and their problem-solving activities

Phases of Data Analytics	Major Types of Insight	Insights Components	Problem-solving Activities
Data Exploration	Analytic insight	• Identification insight	• Search and Filter
		• Perceptive insight	• Perceive and Interpret
Information Synthesis	Synergic insight	• Integrative insight	• Integrate and Synthesize
		• Comprehensive insight	• Connect and Build
Knowledge Actualization	Prognostic insight	• Predictive insight	• Predict and Simulate
		• Prescriptive insight	• Optimize and Assess Risk

This study asserts that to gain actionable insight, the data analytics systems should provide explicit support to data analysts to effectively carry out their problem-solving activities. By supporting these problem-solving activities, the systems enable users to achieve the corresponding insight components that are required to form actionable insight. From this perspective, most of the data analytics systems nowadays only explicitly support the data analysts to achieve analytic insight, leaving them to mentally contend with the processes required to achieve synergic insight and prognostic insight.

8.2.3 Answering Research Question: How can the processes and requirements be effectively supported?

By understanding the leverage points that can be supported to improve user performance in the problem-solving activities, a set of design requirements has been formulated. The review of these design requirements suggests that they require a balanced mixture between: 1) the flexibility of human knowledge and reasoning and 2) the rigor and efficiency of machine-driven computation.

These design requirements are used to develop a conceptual design framework. The framework adopts "machine-augmented cognition" as its design philosophy to guide the design of its solutions for addressing the requirements. The ideology of the *machine-augmented cognition* approach is to amplify the human analytical reasoning capability with computer-aided techniques, with the goal of solving complex analytics problems. In this approach, human analysts prime the data analytic by proactively providing the contextual information and subjective inputs, deciding on and fine-tuning the algorithms, and overwriting the computations with their logic and reasoning. The computational aids are to enhance their analytical reasoning process, rather than automate the analytical reasoning.

The conceptual design framework comprises seven design principles. Each of these is formulated to address a subset of the design requirements. The objective of the design principles is to specify a design of data analytics systems that can effectively support the problem-solving activities in complex analytical tasks. *Table 39* shows the seven design principles.

Table 39. Design principles and problem-solving activities

Phases of Data Analytics	Problem-solving Activities	Design Principles	Result
Data Exploration	• Search and Filter	• Enabling divergent exploration	No
	• Perceive and Interpret	• Enabling managed observations • Enabling exploration convergence	No
Information Synthesis	• Integrate and Synthesize	• Enabling knowledge creation	Yes
	• Connect and Build	• Enabling assisted situation modelling	Yes
Knowledge Actualization	• Predict and Simulate	• Enabling predictive reasoning	Yes
	• Optimize and Assess Risk	• Enabling stochastic optimization	Yes

The design principle "enabling divergent exploration" is to support data analysts to effectively explore a large number of data elements. It includes 1) a design to enable multiple datasets to be integrated to create a centralized data source for generating analytic enquiries and 2) a design to enable analytic enquiries to be generated for perceiving multi-facets of the data. The design principle is formulated to aid the search & filter activity in data analytics.

The design principle "enabling managed observation" is to support data analysts to capture, manage, and retrieve the observations they gain from the analytic enquiries. It includes a design to systematically capture the observations and its components. The design principle is formulated to aid the perceive & interpret activity in data analytics. However, the data analysis suggests that user performance in the activities was not significantly improved.

The design principle "enabling exploration convergence" is to support data analysts to create a joint summary from their observations. It consists of a design to allow the data analysts to visualize and analyze the meta-observations information, and allowing them to make observation about the observations. The design principle is formulated to aid the perceive & interpret activity in data analytics. However, the data analysis suggests that user performance in the activity was not significantly improved.

The design principle "enabling knowledge creation" is to support data analysts to create high-level knowledge by integrating low-level observations and synthesizing them with their implicit knowledge. It consists of 1) a design to integrate multiple observations into high-level knowledge and 2) a design

to structurally capture, store, and retrieve the reasoning used to derive the knowledge. The design principle is formulated to aid the integrate & synthesize activity in data analytics. The data analysis has confirmed that user performance in the activities has been improved.

The design principle "enabling assisted situation modelling" is to support the users to intuitively build a situation model that represents the problem situation as whole. It includes 1) a design to use both quantitative and qualitative information to create the situation model and 2) a design to allow data analysts to visually build the situation model while the underlying inference engine automatically generates the mathematical presentation of the model. The design principle is formulated to aid the connect & build activity in data analytics. The data analysis has confirmed that user performance in the activities has been improved.

The design principle "enabling predictive reasoning" is to enable the users to engage in human steerable prediction & simulation process, aided by computer-aided reasoning techniques. It includes a design of a semi-automatic prediction and simulation engine driven by human reasoning to understand the effects of various assumptions and speculations of the analysts on the situation model. The design principle is formulated to aid the prediction & simulation activity in data analytics. The data analysis has confirmed that user performance in the activities has been improved.

The design principle "enabling stochastic simulation" is to enable the data analysts to optimize the resource allocation for their courses of action, with the goal to meet conflicting objectives within the resource constraints, while compensating for the risk caused by uncertainty. It includes 1) a design to use computational techniques to effectively optimize the resource allocation and 2) a design to allow the data analysts to accurately and rigorously assess the risk associated with the potential courses of action. The design principle is formulated to aid the optimize & assess risk activity in data analytics. The data analysis has confirmed that user performance in the activity has been improved.

In answering the question "how are the processes and requirements can be effectively supported?", empirical findings in this study show that

- the design principle "enabling knowledge creation" can effectively support the integrate & synthesize activity
- the design principle "enabling assisted situation modeling" can effectively support the connect & build activity
- the design principle "enabling predictive reasoning" can effectively support the predict & simulate activity

- the design principle "enabling stochastic optimization" can effectively support the optimize & assess risk activity

However, there is not sufficient evidence to shows that the design principles proposed can support the perceive & interpret activity better than conventional data analytics systems do. This finding is not a major concern because other research and design works have produced alternative designs for supporting the perceive & interpret activity. Although the design principle "enabling managed observations" and "enabling exploration convergence" may not directly support the data exploration, this study asserts that they are important as the enablers for the other design principles to work.

In term of achieving actionable insight, the designs proposed allow the users to achieve high-level insight components, namely synergic insight and prognostic insight. The designs enable the users to comprehend the big picture of the problem situation and to understand the risks associated with their courses of action. In other words, the designs can help to reduce the gap between low-level technical analytic results and the high-level conceptual understandings which are required to solve the analytics problem. Therefore, the users are more likely to achieve actionable insight from the data analytics.

8.3 Contributions

This study provides a number of contributions to the data analytics research community. The contributions range from abstract knowledge to specific design.

8.3.1 A Definition of Actionable Insight

The definition of actionable insight proposed by this study can be useful to spark off intellectual conversations in the data analytics research community. Existing definitions of actionable insight are generally too abstract and ambiguous, which becomes a detriment for the researchers engaged in in-depth discussions. This study discovered that many researchers have used the term in their works, but mostly just mentioned the term in passing, without any further explanation. Some may presume the readers can understand the term as intended by the authors. But this study suggests that, as one of the key terminology in data analytics, the term should be formally and systematically defined, and thus to allow the readers to precisely interpret the research works with minimal ambiguity. The definition of actionable insight in this study is far less than perfect, but this study believes the intellectual conversations in the research community is the best venue for progressively and collectively refine the definition.

Another importance aspect of having a systematic and theoretically-driven definition of actionable insight is to facilitate the design and evaluation of data analytics systems. Such a definition, backed with a full-blown concept, can be useful for informing the developers and researchers about the qualities

their designs seek to achieve, thus allowing them to design systems that can effectively support these qualities. Therefore, an understanding of actionable insight could be the first step towards more effective analytics systems. The definition is also no less important for researchers to develop measurement instruments, allowing them to evaluate whether a data analytics system has lived up to its claims in term of what it is supposed to deliver. This study believes that a measurement instrument developed around action insight would be more meaningful for data analytics research, rather than relying on general measurements such as technology acceptance model (TAM) and usability.

8.3.2 Understanding of Complex Data Analytics Tasks

The conceptual explanatory framework can usefully contribute to the understanding of data analytics involving the complex problem situation. The framework is a result of synthesis between multiple justificatory theories originating in other research domains. These separate pieces of justificatory knowledge are scattered across different sources and are varied in their research contexts. The conceptual explanatory framework, which in this study synthesized and contextualized them specifically in the context of data analytics, can be a native theoretical reference for future data analytics researchers. This study hopes that the framework can save the researchers' time from reinventing the wheel, enabling them to spend valuable time and resources on the novel and value-added aspects of the research, such as design and implementation.

The most important outcome of this study is the overall data analytics workflow and framework designed to support the processes in a complex analytical task. Specifically, this study proposes the idea that a complete data analytics process comprises the data exploration, information synthesis, and knowledge actualization phases. Within each of these phases, specific problem-solving activities need to be performed effectively by data analysts in order to achieve a specific form of knowledge about the analytics problem. The data analysts need to acquire the knowledge from the three phases of data analytics to be able to make informed decision to solve the analytics problem. The overall data analytics workflow and framework are manifested by the explanatory framework.

The explanatory framework provides holistic explanations on data analytics process. It includes the discussions on problem-solving activities, information-processing needs, cognitive states, and the challenges faced by data analysts. The framework can be used by researchers or practitioners to derive design requirements. The framework is still in its infancy and needs to be further refined. As a long-term endeavor, this study hopes that the framework can be a conceptual repository where future findings can be used to reflect upon and improve the framework, not just for the authors of this study. The framework can also be a central repository for the collective knowledge in the data analytics research area, allowing the researchers to communicate their findings using a shared and common basis for discussion.

8.3.3 Design Recommendations for Data Analytics Systems

The conceptual design framework in this study can be a useful reference for informing the design of data analytics systems. In particular, it provides design information on the machine-human corporative approach called "machine-augmented cognition". Significant research works have focused on computation-driven analytics approach, but very few focus on the design of human-driven analytics approaches for complex problem solving. Therefore, this design framework can contribute to this design gap in the data analytics landscape.

The framework provides design recommendations that are grounded in conceptual justifications and empirical validation. It enables practitioners or researchers to determine which of the design recommendations are more suitable for their contexts, allowing them to focus their resource and time on the designs that have empirically proven to be effective.

Moreover, the framework also contains the details on how the conceptual designs can be translated into tangible system functionalities. There is a gap in the data analytics area which occurs because most studies focus on high-level conceptual designs, often leaving too little detail for the practitioners and researchers to realize the design into practical information system artefacts. In contrast, studies that contains low-level details for actualizing the design are rare. In some cases, only the final systems are shown but the rationale and low-level design unpinning the system are unclear. The design framework in this study can be part of the endeavor to reduce the imbalance between abstract studies and low-level designs.

8.3.4 An Implementation Reference

This study also produced a working prototype of data analytics systems. The prototype system is developed as a manifestation of the design framework. It can be helpful as a reference for researchers to have a solid understanding and a better appreciation of the theoretical aspect of this study. Moreover, it can also be a useful and tangible example that helps practitioners to translate the design into system functionalities.

As noted, the primary outcome of this study is the overall workflow and framework of data analytics. The specific computer-aided techniques used in the prototype system are just one way to instantiate the data analytics framework proposed in this study. These computer-aided techniques can be viewed as the swappable modules to fit in data analytics framework, depending on the specific contexts. For instance, in information synthesis phase, the purpose of the design is to support the data analysts to have a dynamic big picture of their problem situation. In the prototype system, Bayesian network modeling is used as the reasoning engine to achieve this purpose. Depending on the need, other computer-aided reasoning techniques such as system dynamics modeling or agent-based modeling can be used to fulfill

the information synthesis described by this proposed data analytics framework. In other words, the workflow and framework for data analytics proposed by this study can act as an integration framework to allow highly technical research works to be applied in a specific application context. *Figure 87* shows how different technical analysis techniques can be used to support the three-phase data analytics framework proposed in this study.

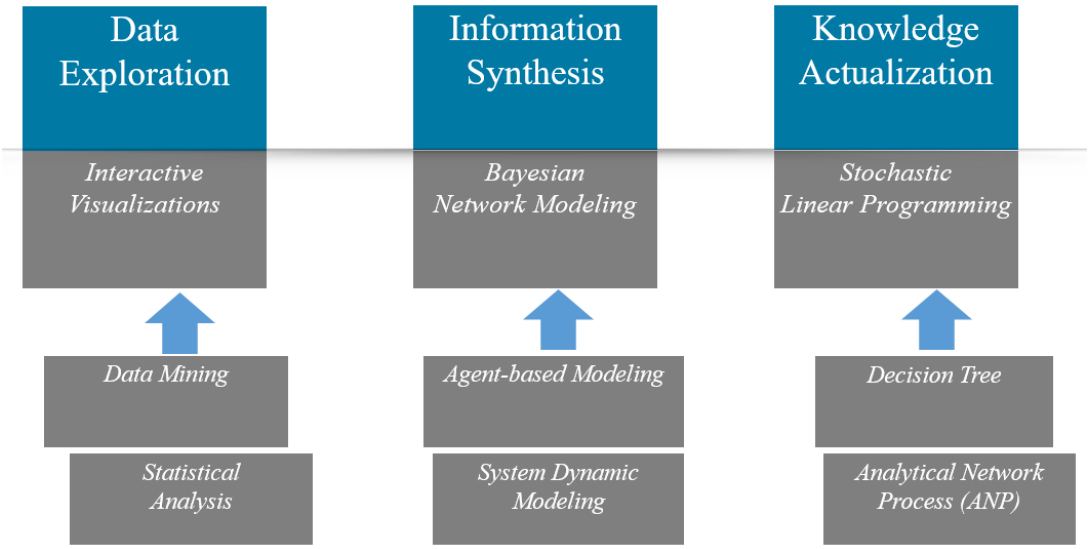


Figure 87. Integrative framework for technical data analysis techniques

One key goal of research is to have practical research outcomes. The prototype is a high-fidelity system that demonstrates value in practice. With the arrangement from QUT Bluebox team, the Capital Market Cooperative Research Center (CMCRC) has had several meetings with the research team to review the funding opportunity in the prototype system. They have pointed out one key feature that attracts their interest: the capability of the system to allow their data analysts to incorporate their accumulated knowledge and experience, in order to have tailored data analytics.

8.4 Limitations and Future Works

The explanatory framework and design framework are developed to target the complex problem situation. Thus, the designs which are derived based on the requirements may also not be applicable to other types of data analytics, particular a very well-structured problem. Moreover, the problem-solving activities in the complex analytical task may be different in other types of data analytics. Future studies can develop a variation of the explanatory framework and design framework to suit other types of data analytics. Both the explanatory and design frameworks can be used as the general classes of which its sub-elements and features can be adapted to suit the specific context of the future study. For instance, data analysts in a crime & intelligent analysis similarly follow through the three-phase process of data exploration, information synthesis, and knowledge actualization, but the specific problem-solving activities during the knowledge actualization phase may be different. The analysts may not need to

optimize the resource allocation to fulfill multiple conflicting objectives and constraints. Instead, they might need to weight alternative hypotheses based on the evidence associated with the hypotheses.

Although the design principles are general and can be apply to different complex analytics problems, the prototype instantiated the design principles in the stock investment analysis domain. The result of the data analysis is based on evaluation of the prototype. Therefore, the findings of how the design could affect the analysts' analytical performance is validated only in the stock analysis domain. Caution should be taken when generalize the findings to other domains. Future research can replicate the study in other analytics problem areas such as anti-terrorism analysis or environment policy development to cross-validate the findings in this study.

More importantly, the due to its broad research scope, the purpose of this study is to introduce a data analytics framework that covers data exploration, information synthesis, and knowledge actualization. However, each of these phases in data analytics is large enough to be a research study on its own. This study covers only sufficient depth for the purpose of discussing and meeting the research objectives. Future studies can specifically focus on a particular phase of the data analytics to achieve much deeper research investigation. For instance, future studies can exclusively focus on information synthesis phase to carry out in-depth studies that include expert interview, theory validation, prototype evaluation, and user interview. These future works can refine the existing framework to become a well-validated and robust framework that can be readily use by designers and developers for building effective data analytics systems.

In terms of data collection and analysis, this study was not able to examine the isolated effects of each design principle. Due to the research scope, this study has too many variables to be controlled in order to study the design effects in isolation. To achieve that requires significant resource and time to conduct many different combinations of the settings. Such arrangement is impractical given the constrains in resource and time in Ph.D. studies. Future studies can be designed to specifically study the impacts of the design principle on the user behaviors. Such studies can focus on a single design principle and a limited set of functionalities at a time. Due to the conceptual nature of a design principle, it can be realized in different designs. Future studies that focus on single design principle can produce in-depth research findings that give deeper insight into how a specific problem-solving activity can be better supported. More importantly, such in-depth studies can afford to compare multiple techniques, in order to identify the technique that can address the design principle most effectively. Due to the well-defined scope and objectives of these studies, it is possible to use eye-tracking device and Electroencephalogram (EEG) reader to scrutinize the impact of the design.

8.5 Final Remarks

The findings of this study contribute to a better understanding of what is an actionable insight, how it can be achieved, and how data analytics systems can be better designed to help users to achieve actionable insight. The systematic and theory-grounded definition and process of actionable insight proposed by this study can be useful for researchers to develop alternative theories on the process of actionable insight and to develop robust measurement instruments for evaluating data analytics systems. The design framework can be used by practitioners such as software developers to design data analytics systems that provide effective supports for all the analyst's problem-solving activities. As for the policy makers, the findings of this study have shown that it is more important to have users who have rich domain knowledge and experience than to invest in data analytics systems that too advanced for the users to use.

As data becomes a necessity for modern institutions to survive and thrive, data analytics systems are becoming ever more important for these institutions to convert the investments in the data into real value. This real value can be achieved only if the data analytics can be used to intuitively inform decision and devise a value-added action plan. To achieve this, the data analysts are inseparable from the data analytics. It is critical that the data analysts are supported to effectively perform the high-level analytic activities. A human-computer symbiosis approach is required to meet this requirement. This study contributes to the design of data analytics systems that can effectively support the high-level analytic activities required to achieve actionable insight. It refocuses the role of data analytics as being the most critical part of the data analytics.

Based on this notion, this study believes that the advancements in automated data analytics such as machine learning and artificial intelligence do not replace the data analysts, but such advancements up-shift the work load of the data analysts to focus on the higher-level data analytics activities that are even more intellectually challenging. The abundance of information may have led humans to consider it to be as boring as data is considered today. This upward shift allows humans to find greater interest in knowledge creation, scenario building, and solution planning that is directed towards turning vision into reality.

Reference

Albers, M. J. (1999). *Information design considerations for improving situation awareness in complex problem-solving*. Paper presented at the Proceedings of the 17th annual international conference on Computer documentation, New Orleans, Louisiana, USA.

Amar, R. A., & Stasko, J. (2004). *A Knowledge Task-Based Framework for Design and Evaluation of Information Visualizations*. Paper presented at the Proceedings of the IEEE Symposium on Information Visualization.

Amar, R. A., & Stasko, J. T. (2005). Knowledge Precepts for Design and Evaluation of Information Visualizations. *IEEE Transactions on Visualization and Computer Graphics, 11*(4), 432-442.

Arazy, O., Kumar, N., & Shapira, B. (2010). A Theory-Driven Design Framework for Social Recommender Systems. *Journal of the Association for Information Systems, 11*(9), 455.

Attfield, S., Hara, S., & Wong, B. (2010). Sense-making in visual analytics: processes and challenges.

Ayres, P., & Van Gog, T. (2009). State of the art research into cognitive load theory. *Computers in Human Behavior, 25*(2), 253-257.

Basole, R. C., Hu, M. D., Patel, P., & Stasko, J. T. (2012). Visual Analytics for Converging-Business-Ecosystem Intelligence. *Ieee Computer Graphics and Applications, 32*(1), 92-96.

Beath, C., Irma, B.-F., Ross, J., & Short, J. (2012). Finding Value in the Information Explosion. *MIT Sloan Management Review*.

Bertini, E., & Lalanne, D. (2009). *Surveying the complementary role of automatic data analysis and visualization in knowledge discovery*. Paper presented at the Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery: Integrating Automated Analysis with Interactive Exploration, Paris, France.

Bose, R. (2009). Advanced analytics: opportunities and challenges. *Industrial Management & Data Systems, 109*(2), 155-172.

Burby, J., & Atchison, S. (2007). *Actionable Web Analytics : Using Data to Make Smart Business Decisions*

Calisir, F., & Calisir, F. (2004). The relation of interface usability characteristics, perceived usefulness, and perceived ease of use to end-user satisfaction with enterprise resource planning (ERP) systems. *Computers in Human Behavior, 20*(4), 505-515.

Cao, L. (2012). Actionable knowledge discovery and delivery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2*(2), 149-163.

Cao, L., Luo, D., & Zhang, C. (2007). Knowledge actionability: satisfying technical and business interestingness. *Int. J. Bus. Intell. Data Min., 2*(4), 496-514.

Carlsson, S. A. (2010). Design Science Research in Information Systems: A Critical Realist Approach *Design Research in Information Systems: Theory and Practice* (pp. 209-233). Boston, MA: Springer US.

Carpi, A., & Egger, A. (2008). Experimentation in Scientific Research *Vision Learning, 1*(7). Retrieved from http://www.visionlearning.com/en/library/Process-of-Science/49/Experimentation-in-Scientific-Research/150

Chabot, C. (2009). Demystifying Visual Analytics. *Ieee Computer Graphics and Applications, 29*(2), 84-87.

Chang, R., Ziemkiewicz, C., Green, T. M., & Ribarsky, W. (2009). Defining Insight for Visual Analytics. *Computer Graphics and Applications, IEEE, 29*(2), 14-17.

Chen, J. Q., & Lee, S. M. (2003). An exploratory cognitive DSS for strategic decision making. *Decision Support Systems, 36*(2), 147-160.

Chinchor, N., & Pike, W. A. (2009). The Science of Analytic Reporting. *Information Visualization, 8*(4), 286-293.

Chiu, S., & Tavella, D. (2008). *Data Mining and Market Intelligence for Optimal Marketing Returns*. San Francisco: Taylor & Francis.

Courtney, J. F. (2001). Decision making and knowledge management in inquiring organizations: toward a new decision-making paradigm for DSS. *Decision Support Systems, 31*(1), 17-38.

Crandall, B., Klein, G., & Hoffman, R. R. (2006). *Working minds a practitioner's guide to cognitive task analysis*. London, England: Cambridge, Massachusetts.

Cross, R., & Sproull, L. (2004). More Than an Answer: Information Relationships for Actionable Knowledge. *Organization Science, 15*(4), 446-462.

David, G., & Michelle, X. Z. (2009). Characterizing users' visual analytic activity for insight provenance. *Information Visualization, 8*(1), 42-55.

Demirer, R., Mau, R. R., & Shenoy, C. (2006). Bayesian Networks: A Decision Tool to Improve Portfolio Risk Analysis. *Journal of Applied Finance, 16*(2), 106-119.

Di, Y., Rundensteiner, E. A., & Ward, M. O. (2007, Oct. 30 2007-Nov. 1 2007). *Analysis Guided Visual Exploration of Multivariate Data.* Paper presented at the Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on.

Eckerson, W. W. (2009). *Beyond reporting: Deliverying insights with next-generation analytics*.

Eick, S. G. (2000). Visual Discovery and Analysis. *IEEE Transactions on Visualization and Computer Graphics, 6*(1), 44-58.

Endsley, M. R. (1995a). Measurement of Situation Awareness in Dynamic Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 37*(1), 65-84.

Endsley, M. R. (1995b). Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 37*(1), 32-64.

Endsley, M. R., Bolte, B., & Jones, D. G. (2011). *Designing for Situation Awareness: An Approach for User-Centered Design*. Georgia, USA: Taylor & Francis

Endsley, M. R., & Garland, D. (2000). Theoretical underpinnings of situation awareness: A critical review. *Situation awareness analysis and measurement*, 3-32.

Endsley, M. R., & Jones, D. G. (2011). *Designing for Situation Awareness An Approach to User-Centered Design* (Second Edition ed.): CRC Press 2011.

Endsley, M. R., Selcon, S. J., Hardiman, T. D., & Croft, D. G. (1998). A Comparative Analysis of Sagat and Sart for Evaluations of Situation Awareness. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 42*(1), 82-86. doi:10.1177/154193129804200119

Eppler, M. J., & Platts, K. W. (2009). Visual Strategizing: The Systematic Use of Visualization in the Strategic-Planning Process. *Long Range Planning, 42*(1), 42-74.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery: an overview. In M. F. Usama, P.-S. Gregory, S. Padhraic, & U. Ramasamy (Eds.), *Advances in knowledge discovery and data mining* (pp. 1-34): American Association for Artificial Intelligence.

Feng, Y.-H., Teng, T.-H., & Tan, A.-H. (2009). Modelling situation awareness for Context-aware Decision Support. *Expert Systems with Applications, 36*(1), 455-463.

Funke, J. (2010). Complex problem solving: a case for complex cognition? *Cognitive Processing, 11*(2), 133-142.

Garg, S., Nam, J. E., Ramakrishnan, I. V., & Mueller, K. (2008). Model-Driven Visual Analytics. *Ieee Symposium on Visual Analytics Science and Technology 2008, Proceedings*, 19-26.

Gary, M. S., & Wood, R. E. (2011). Mental models, decision rules, and performance heterogeneity. *Strategic management journal, 32*(6, (6)), 569-594.

Glykas, M. (Ed.) (2010). *Fuzzy Cognitive Maps: Advances in Theory, Methdologies, Tools and Applications* (Vol. 247): Springer Berlin Heidelberg.

Gore, J., Banks, A., Millward, L., & Kyriakidou, O. (2006). Naturalistic Decision Making and Organizations: Reviewing Pragmatic Science. *Organization Studies, 27*(7), 925-942.

Gotz, D., When, Z., Lu, J., Kissa, P., Cao, N., Qian, W. H., . . . Zhou, M. X. (2010). *HARVEST: an intelligent visual analytic tool for the masses*. Paper presented at the Proceedings of the first international workshop on Intelligent visual interfaces for text analysis, Hong Kong, China.

Gotz, D., & Zhou, M. X. (2008). *Empirical study of user interaction behavior during visual analytics*. Retrieved from

Gotz, D., Zhou, M. X., & Aggarwal, V. (2006, Oct. 31 2006-Nov. 2 2006). *Interactive Visual Synthesis of Analytic Knowledge*. Paper presented at the Visual Analytics Science And Technology, 2006 IEEE Symposium On.

Green, T. M., & Maciejewski, R. (2013). *A role for reasoning in visual analytics.* Paper presented at the System Sciences (HICSS), 2013 46th Hawaii International Conference on.

Green, T. M., Ribarsky, W., & Fisher, B. (2008, 19-24 Oct. 2008). *Visual analytics for complex concepts using a human cognition model.* Paper presented at the Visual Analytics Science and Technology, 2008. VAST '08. IEEE Symposium on.

Green, T. M., Ribarsky, W., & Fisher, B. (2009). Building and applying a human cognition model for visual analytics. *Information Visualization, 8*(1), 1-13.

Green, T. M., Wakkary, R., Arias, H., x, & ndez, R. (2011, 4-7 Jan. 2011). *Expanding the Scope: Interaction Design Perspectives for Visual Analytics.* Paper presented at the System Sciences (HICSS), 2011 44th Hawaii International Conference on.

Gregor, S., & Jones, D. (2007). The Anatomy of a Design Theory. *Journal of the Association for Information Systems, 8*(5), 312-323,325-335.

Greiff, S. (2012). Assessment and Theory in Complex Problem Solving - A Continuing Contradiction? *Journal of Educational and Developmental Psychology, 2*(1), 49-56.

Groth, D. P., & Streefkerk, K. (2006). Provenance and Annotation for Visual Exploration Systems. *Visualization and Computer Graphics, IEEE Transactions on, 12*(6), 1500-1510.

Hall, R. (1998). Within-Subjects Designs. *Psychology World.* Retrieved from https://web.mst.edu/~psyworld/experimental/within_subjects.html

Ham, D. H. (2010). The State of the Art of Visual Analytics. *Ekc 2009 Proceedings of Eu-Korea Conference on Science and Technology, 135*, 213-222.

Harris, J. G. (2005). *The Insight-to-Action Loop - Transforming information into business performance*. Retrieved from

Haynie, J. M., Shepherd, D., Mosakowski, E., & Earley, P. C. (2010). A situated metacognitive model of the entrepreneurial mindset. *Journal of Business Venturing, 25*(2), 217-229. doi:http://dx.doi.org/10.1016/j.jbusvent.2008.10.001

Heer, J., Mackinlay, J., Stolte, C., & Agrawala, M. (2008). Graphical Histories for Visualization: Supporting Analysis, Communication, and Evaluation. *IEEE Transactions on Visualization and Computer Graphics, 14*(6), 1189-1196. doi:10.1109/tvcg.2008.137

Heer, J., & Shneiderman, B. (2012a). Interactive dynamics for visual analysis. *Queue, 10*(2), 30.

Hetzler, B., Whitney, P., Martucci, L., & Thomas, J. (1998, 19-20 Oct 1998). *Multi-faceted insight through interoperable visual information analysis paradigms.* Paper presented at the Information Visualization, 1998. Proceedings. IEEE Symposium on.

Heuer, R. J. (1999). *Psychology of Intelligencen Analysis*. United States: Center for the Study of Intelligence.

Houxing, Y. (2010, 22-23 May 2010). *A Knowledge Management Approach for Real-Time Business Intelligence.* Paper presented at the 2010 2nd International Workshop on Intelligent Systems and Applications (ISA).

Hui, R. (2014). *Measuring the Quality of Actionable Insight.* (Honors), Queensland University of Technology.

IBM Global Business Service. (2010). *Analytics: The new path to value*.

IDC. (2014). The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things. *EMC Digital Universe with Research & Analysis*. Retrieved from http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm

Jankun-Kelly, T. J., Kwan-Liu, M., & Gertz, M. (2007). A Model and Framework for Visualization Exploration. *Visualization and Computer Graphics, IEEE Transactions on, 13*(2), 357-369. doi:10.1109/tvcg.2007.28

Jarzabkowski, P., & Wilson, D. C. (2006). Actionable Strategy Knowledge:: A Practice Perspective. *European Management Journal, 24*(5), 348-367.

Jonassen, D. H. (2000). Toward a design theory of problem solving. *Educational Technology Research and Development, 48*(4), 63-85.

Jonassen, D. H. (2012). Designing for decision making. *Educational Technology, Research and Development, 60*(2), 341-359.

Kandel, S., Paepcke, A., Hellerstein, J., & Heer, J. (2012). *Enterprise Data Analysis and Visualization: An Interview Study*. Paper presented at the IEEE Visual Analytics Science & Technology (VAST).

Keim, D. A. (2002). Information Visualization and Visual Data Mining. *IEEE Transactions on Visualization and Computer Graphics, 8*(1), 1-8. doi:10.1109/2945.981847

Keim, D. A. (2012). Solving Problems with Visual Analytics: Challenges and Applications. In P. Flach, T. Bie, & N. Cristianini (Eds.), *Machine Learning and Knowledge Discovery in Databases* (Vol. 7523, pp. 5-6): Springer Berlin Heidelberg.

Keim, D. A., Kohlhammer, J., Ellis, G., & Mansmann, F. (Eds.). (2010). *Solving Problems with Visual Analytics*. Goslar, Germany: Eurographics Association.

Keim, D. A., Mansmann, F., Schneidewind, J., Thomas, J., & Ziegler, H. (2008). Visual Analytics: Scope and Challenges. In S. Simoff, M. Böhlen, & A. Mazeika (Eds.), *Visual Data Mining* (Vol. 4404, pp. 76-90): Springer Berlin Heidelberg.

Keim, D. A., Mansmann, F., Schneidewind, J., & Ziegler, H. (2006). *Challenges in Visual Data Analysis*. Paper presented at the Proceedings of the conference on Information Visualization.

Kirsh, D. (2009). Problem Solving and Situated Cognition. In P. Robbins & M. Aydede (Eds.), *The Cambrige Handbook of Situated Cognition* (pp. 264-306). Cambrige: Cambrige University Press.

Kirsh, D. (2010). Thinking with external representations. *AI & Society, 25*(4), 441-454.

Klein, G. (1993). *Naturalistic decision making: Implications for design*. Retrieved from

Klein, G. (1997). The recognition-primed decision (RPD) model: Looking back, looking forward. In C. E. Z. G. Klein (Ed.), *Naturalistic decision making* (pp. 285-292). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Klein, G. (1999). *Sources of Power: How People Make Decisions*: The MIT Press.

Klein, G., Moon, B., & Hoffman, R. R. (2006a). Making Sense of Sensemaking 1: Alternative Perspectives. *Intelligent Systems, IEEE, 21*(4), 70-73.

Klein, G., Moon, B., & Hoffman, R. R. (2006b). Making Sense of Sensemaking 2: A Macrocognitive Model. *Intelligent Systems, IEEE, 21*(5), 88-92.

Kohlhammer, J., Keim, D. A., Pohl, M., Santucci, G., & Andrienko, G. (2011). Solving Problems with Visual Analytics. *Proceedings of the 2nd European Future Technologies Conference and Exhibition 2011 (Fet 11), 7*, 117-120.

Kohlhammer, J., May, T., & Hoffmann, M. (2009). Visual Analytics for the Strategic Decision Making Process. In R. Amicis, R. Stojanovic, & G. Conti (Eds.), *Geospatial Visual Analytics* (pp. 299-310): Springer Netherlands.

Kokar, M. M., & Endsley, M. R. (2012). Situation Awareness and Cognitive Modeling. *Intelligent Systems, IEEE, 27*(3), 91-96. doi:10.1109/MIS.2012.61

Kuechler, W., & Vaishnavi, V. (2012). A Framework for Theory Development in Design Science Research: Multiple Perspectives. *Journal of the Association for Information Systems, 13*(6), 395-423.

Lee, K. J., & Chang, W. (2009). Bayesian belief network for box-office performance: A case study on Korean movies. *Expert Systems with Applications, 36*(1), 280-291.

Lee, S. M., & Chen, Q. (1997). A conceptual model for executive support systems. *Logistics Information Management, 10*(4), 154-159.

Lefebvre, S. (2004). A Look at Intelligence Analysis. *International Journal of Intelligence and CounterIntelligence, 17*(2), 231-264.

Leppink, J., Paas, F., Van der Vleuten, C. P., Van Gog, T., & Van Merriënboer, J. J. (2013). Development of an instrument for measuring different types of cognitive load. *Behavior research methods, 45*(4), 1058-1072.

Liebowitz, J. (2013). *Big data and business analytics*: CRC Press.

Ling, X., Gerth, J., & Hanrahan, P. (2006, Oct. 31 2006-Nov. 2 2006). *Enhancing Visual Analysis of Network Traffic Using a Knowledge Representation.* Paper presented at the Visual Analytics Science And Technology, 2006 IEEE Symposium On.

Lipford, H. R., Stukes, F., Wenwen, D., Hawkins, M. E., & Chang, R. (2010, 25-26 Oct. 2010). *Helping users recall their reasoning process.* Paper presented at the Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on.

Lu, J., Niu, L., & Zhang, G. (2012). A Situation Retrieval Model for Cognitive Decision Support in Digital Business Ecosystems. *Industrial Electronics, IEEE Transactions on, PP*(99), 1-1.

Lund, A. M. (2001). Measuring Usability with the USE Questionnaire12.

Markus, M. L., Majchrzak, A., & Gasser, L. (2002). A design theory for systems that support emergent knowledge processes. *MIS Q., 26*(3), 179-212.

Meso, P., Troutt, M. D., & Rudnicka, J. (2002). A review of naturalistic decision making research with some implications for knowledge management. *Journal of Knowledge Management, 6*(1), 63-73.

Meyer, J., Thomas, J., Diehl, S., Fisher, B., & Keim, D. A. (2010). From visualization to visually enabled reasoning. *Dagstuhl Follow-Ups, 1*.

Mills, J. H., Thurlow, A., & Mills, A. J. (2010). Making sense of sensemaking: the critical sensemaking approach. *Qualitative Research in Organizations and Management: An International Journal, 5*(2), 182-195.

Mintzberg, H., & Westley, F. (2001). Decision Making: It's not what you think. *MIT Sloan Management Review, 42*(3), 89-93.

Mirel, B. (2001). Testing the usability of interactive visualizations for complex problem-solving: Findings related to improving interfaces and help. *Journal of Technical Writing and Communication, 31*(1), 7-26

Mirel, B. (2004). *Interaction design for complex problem solving: Developing useful and usable software*. San Francisco, U.S.A: Elsevier.

Mirel, B., & Allmendinger, L. (2004). Visualizing complexity: Getting from here to there in ill-defined problem landscapes. *Information Design Journal, 12*(2), 141-151.

Nakatsu, R. (2010). *Diagrammatic Reasoning in AI*. United States: John Wiley & Sons

Nemati, H., Earle, B., Arekapudi, S., & Mamani, S. (2010). Do users go both ways?: BI user profiles fit BI tools. *International Journal of Business Intelligence Research, 1*(3), 15-33. doi:10.4018/jbir.2010070102

Niu, L., Lu, J., & Zhang, G. (2009). Cognition-driven decision support for business intelligence. *Models, Techniques, Systems and Applications. Studies in Computational Intelligence, Springer, Berlin.*

Niu, L., Lu, J., & Zhang, G. (2009). *Cognition-Driven Decision Support for Business Intelligence: Models, Techniques, Systems and Applications*. Chennai, India: Springer.

Ntuen, C. A. (2009). *Sensemaking as a naturalistic knowledge discovery model.* Paper presented at the Naturalistic Decision Making and Computers Conference, London,(June).

Ntuen, C. A., Park, E. H., & Gwang-Myung, K. (2010). Designing an Information Visualization Tool for Sensemaking. *International Journal of Human-Computer Interaction, 26*(2-3), 189-205.

Nwiabu, N., Allison, I., Holt, P., Lowit, P., & Oyeneyin, B. (2011, 22-24 Feb. 2011). *Situation awareness in context-aware case-based decision support.* Paper presented at the Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), 2011 IEEE First International Multi-Disciplinary Conference on.

Ohsawa, Y., & Nishihara, Y. (2012). Chance Discovery as Value Sensing for Innovation *Innovators' Marketplace* (pp. 15-31): Springer.

Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. M. (2003). Cognitive Load Measurement as a Means to Advance Cognitive Load Theory. *Educational Psychologist, 38*(1), 63-71.

Paas, F., Van Gog, T., & Sweller, J. (2010). Cognitive load theory: New conceptualizations, specifications, and integrated research perspectives. *Educational Psychology Review, 22*(2), 115-121.

Parrish, J. (2008). *Sensemaking in information systems: Toward a sensemaking inquiring system.* (Doctor of Philosophy), University of Central Florida, Florida, United States.

Pavel, N., & Dragos, S. (2010). A new business dimension - business analystics *Accounting and Management Information Systems, 9*(4), 603-618.

Pike, W. A., May, R., Baddeley, B., Riensche, R., Bruce, J., & Younkin, K. (2007, 25-25 May 2007). *Scalable visual reasoning: Supporting collaboration through distributed analysis.* Paper presented at the Collaborative Technologies and Systems, 2007. CTS 2007. International Symposium on.

Pike, W. A., Stasko, J., Chang, R., & O'Connell, T. A. (2009). The science of interaction. *Information Visualization, 8*(4), 263-274. doi:10.1057/ivs.2009.22

Pirolli, P., & Card, S. (2005, 2005). *The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis*.

Pohl, M., Smuc, M., & Mayr, E. (2012). The User Puzzle - Explaining the Interaction with Visual Analytics Systems. *Visualization and Computer Graphics, IEEE Transactions on, 18*(12).

Qu, Y., & Furnas, G. W. (2005). *Sources of structure in sensemaking*. Paper presented at the CHI '05 extended abstracts on Human factors in computing systems, Portland, OR, USA.

Ribarsky, W., Fisher, B., & Pottenger, W. M. (2009). Science of Analytical Reasoning. *Information Visualization, 8*(4), 254-262. doi:10.1057/ivs.2009.28

Richmond, B., & Peterson, S. (2001). *An introduction to systems thinking*: High Performance Systems., Incorporated.

Robinson, A. C. (2008, 19-24 Oct. 2008). *Collaborative synthesis of visual analytic results*. Paper presented at the Visual Analytics Science and Technology, 2008. VAST '08. IEEE Symposium on.

Robinson, A. C. (2009). Needs Assessment for the Design of Information Synthesis Visual Analytics Tools. *Information Visualization, Iv 2009, Proceedings*, 353-360. doi:Doi 10.1109/Iv.2009.85

Rudolph, J. W. (2003). Into the big muddy and out again: Error persistence and crisis management in the operating room. *Boston College Dissertations and Theses*, AAI3103269.

Russell, D. M., Stefik, M. J., Pirolli, P., & Card, S. K. (1993). *The cost structure of sensemaking*. Paper presented at the Proceedings of the INTERACT '93 and CHI '93 conference on Human factors in computing systems, Amsterdam, The Netherlands.

Salmon, P. M., Stanton, N. A., Walker, G. H., Jenkins, D., Ladva, D., Rafferty, L., & Young, M. (2009). Measuring Situation Awareness in complex systems: Comparison of measures study. *International Journal of Industrial Ergonomics, 39*(3), 490-500. doi:http://dx.doi.org/10.1016/j.ergon.2008.10.010

Sammon, D. (2008). Understanding Sense-Making *Encyclopedia of Decision Making and Decision Support Technologies* (pp. 916-921): IGI Global.

Sankar, S. (2013). The Risk of Human-Computer Cooperation. Retrieved from https://scn.sap.com/thread/3394408

Saraiya, P., North, C., & Duca, K. (2004, 0-0 0). *An Evaluation of Microarray Visualization Tools for Biological Insight.* Paper presented at the Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on.

Saraiya, P., North, C., & Duca, K. (2005). An insight-based methodology for evaluating bioinformatics visualizations. *Visualization and Computer Graphics, IEEE Transactions on, 11*(4), 443-456.

Saraiya, P., North, C., Lam, V., & Duca, K. A. (2006). An Insight-Based Longitudinal Study of Visual Analytics. *Visualization and Computer Graphics, IEEE Transactions on, 12*(6), 1511-1522.

Sawyer, R. (2011). BI's Impact on Analyses and Decision Making Depends on the Development of Less Complex Applications. *International Journal of Business Intelligence Research, 2*(3), 52-63.

Schneider, R. D., & Gibson, D. (2011). *Microsoft SQL Server 2008 All-in-One Desk Reference For Dummies*: John Wiley & Sons.

Sekaran, U., & Bougie, R. (2009). *Research Methods for Business: A Skill-Building Approach* (5th ed.). United Kingdom: John Wiley & Sons Ltd.

Sell, D., Silva, D. C. d., Beppler, F. D., Napoli, M., Ghisi, F. B., Pacheco, R. C. S., . . . Todesco, L. (2008). *SBI: a semantic framework to support business intelligence*. Paper presented at the Proceedings of the first international workshop on Ontology-supported business intelligence, Karlsruhe, Germany.

Shattuck, L. G., & Miller, N. L. (2006). Extending Naturalistic Decision Making to Complex Organizations: A Dynamic Model of Situated Cognition. *Organization Studies, 27*(7), 989-1009. doi:10.1177/0170840606065706

Shenoy, P. P., & Shenoy, C. (2000). Bayesian network models of portfolio risk and return.

Shrinivasan, Y. B. (2010). *Supporting the Sensemaking Process in Visual Analytics.* (Doctor of Philosophy), Eindhoven University of Technology.

Shrinivasan, Y. B., Gotz, D., & Jie, L. (2009, 12-13 Oct. 2009). *Connecting the dots in visual analysis.* Paper presented at the Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on.

Shrinivasan, Y. B., & Wijk, J. J. v. (2008). *Supporting the analytical reasoning process in information visualization.* Paper presented at the Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems, Florence, Italy.

Siemens, G. (2011). *Orientation: sensemaking and wayfinding in complex distributed online information environments.* (Doctor of Philosophy), University of Aberdeen.

Smuc, M., Mayr, E., Lammarsch, T., Bertone, A., Aigner, W., Risku, H., & Miksch, S. (2008). Visualizations at First Sight: Do Insights Require Training?

HCI and Usability for Education and Work. In A. Holzinger (Ed.), (Vol. 5298, pp. 261-280): Springer Berlin / Heidelberg.

Stanners, M., & French, H. T. (2005). *An empirical study of the relationship between situation awareness and decision making*. Retrieved from

Stijn, V., & Annabel Van den, B. (2011). The Secrets to Managing Business Analytics Projects. *MIT Sloan Management Review, 53*(1), 65-69.

Stubbs, E. (2011). *Value of Business Analytics : Identifying the Path to Profitability*. Hoboken NJ, USA: Wiley.

Surma, J. (2011). *Business Intelligence: Making Decisions Through Data Analytics* (1th ed.). New York, United States: Business Expert Press, LLC.

T., L., Schreck, T., Fellner, D. W., & Kohlhammer, J. (2012). Visual Search and Analysis in Complex Informaiton Spaces - Approaches and Research Challenges. In J. Dill, R. Earnshaw, & D. Kasik (Eds.), *Expanding the Frontiers of Visual Analytics and Visualization* (pp. 45-67). New York: Springer.

Thomas, J., & Cook, K. A. (2005). *Illuminating the Path: The Research and Development Agenda for Visual Analytics* National Visualization and Analytics Centre.

Thomas, J., & Kielman, J. (2009). Challenges for Visual Analytics. *Information Visualization, 8*(4), 309-314. doi:10.1057/ivs.2009.26

Thomas, J. B., Clark, S. M., & Gioia, D. A. (1993). Strategic Sensemaking and Organizational Performance: Linkages among Scanning, Interpretation, Action, and Outcomes. *The Academy of Management Journal, 36*(2), 239-270.

Tim, C. (2006). Enhancing insight discovery by balancing the focus of analytics between strategic and tactical levels. *Journal of Database Marketing & Customer Strategy Management, 13*(4), 261-270.

Tsai, H.-T., Chien, J.-L., & Tsai, M.-T. (2014). The influences of system usability and user satisfaction on continued Internet banking services usage intention: empirical evidence from Taiwan. *Electronic Commerce Research, 14*(2), 137-169. doi:10.1007/s10660-014-9136-5

van Merriënboer, J. J. G., & Sluijsmans, D. M. A. (2009). Toward a Synthesis of Cognitive Load Theory, Four-Component Instructional Design, and Self-Directed Learning. *Educational Psychology Review, 21*(1), 55-66.

Vandewaetere, M., & Clarebout, G. (2013). Cognitive load of learner control: extraneous or germane load? *Education Research International, 2013*.

Walls, J. G., Widmeyer, G. R., & El Sawy, O. A. (1992). Building an Information System Design Theory for Vigilant EIS. *Information Systems Research, 3*(1), 36-59.

Watson, H. J. (2011). Business Analytics Insight: Hype or Here to Stay? *Business Intelligence Journal, 16*, 4-8.

Weick, K. E. (1993). The Collapse of Sensemaking in Organizations: The Mann Gulch Disaster. *Administrative Science Quarterly, 38*(4), 628-652.

Weick, K. E. (1995). *Sensemaking in organizations*. Thousand Oaks: Sage Publications.

Weick, K. E., Sutcliffe, K. M., & Obstfeld, D. (2005). Organizing and the Process of Sensemaking. *Organization Science, 16*(4), 409-421. doi:10.1287/orsc.1050.0133

William, P., Richard, M., & Alan, T. (2007, Jan. 2007). *Supporting Knowledge Transfer through Decomposable Reasoning Artifacts.* Paper presented at the System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on.

Wright, K. (2005). Personal knowledge management: supporting individual knowledge worker performance. *Knowledge Management Research & Practice, 3*(3), 156-165.

Wright, W., Schroh, D., Proulx, P., Skaburskis, A., & Cort, B. (2006). *The Sandbox for analysis: concepts and methods*. Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Montréal, Québec, Canada.

Yadav, S. B., & Khazanchi, D. (1992). Subjective understanding in strategic decision making: An information systems perspective. *Decision Support Systems, 8*(1), 55-71. doi:http://dx.doi.org/10.1016/0167-9236(92)90037-P

Yang, C., Jing, Y., & Ribarsky, W. (2009, 20-23 April 2009). *Toward effective insight management in visual analytics systems.* Paper presented at the Visualization Symposium, 2009. PacificVis '09. IEEE Pacific.

Yi, J. S., Kang, Y.-a., Stasko, J. T., & Jacko, J. A. (2008). *Understanding and characterizing insights: how do people gain insights using information visualization?* Paper presented at the Proceedings of the 2008 Workshop on BEyond time and errors: novel evaLuation methods for Information Visualization, Florence, Italy.

Zhang, P., Soergel, D., Klavans, J. L., & Oard, D. W. (2008). Extending sense-making models with ideas from cognition and learning theories. *Proceedings of the American Society for Information Science and Technology, 45*(1), 23-23. doi:10.1002/meet.2008.1450450219

Zhicheng, L., Nersessian, N. J., & Stasko, J. T. (2008). Distributed Cognition as a Theoretical Framework for Information Visualization. *Visualization and Computer Graphics, IEEE Transactions on, 14*(6), 1173-1180. doi:10.1109/TVCG.2008.121

Zuk, T., & Carpendale, S. (2007). Visualization of Uncertainty and Reasoning

Smart Graphics. In A. Butz, B. Fisher, A. Krüger, P. Olivier, & S. Owada (Eds.), (Vol. 4569, pp. 164-177): Springer Berlin / Heidelberg.

Appendices

10.1 Appendix A – Questionnaire Items

Queensland University of Technology
Brisbane Australia

PARTICIPANT INFORMATION FOR QUT RESEARCH PROJECT
– Questionnaire –

Deriving Actionable Insight: A Problem-solving Approach to Data Analytics

QUT Ethics Approval Number 1600000553

RESEARCH TEAM

Principal Researcher: Shiang Yen Tan, PhD Student, QUT
Associate Researcher: Dr. Taizan Chan, Supervisor
School of Information Systems
Science and Engineering Faculty, Queensland University of Technology (QUT)
Email: t.chan@qut.edu.au

Dr. Ernest Foo, Associate Supervisor
School of Electrical Engineering and Computer Science
Science and Engineering Faculty, Queensland University of Technology (QUT)
Email: e.foo@qut.edu.au

Associate Professor Yue Xu
Email: yue.xu@qut.edu.au

DESCRIPTION

This project is being undertaken as part of PhD study for Shiang Yen, Tan.

The purpose of this project is to develop a design of data analytics systems that can help data analysts to make informed decision based on complex and large datasets. This project has proposed a problem-solving approach to data analytics. A prototype has been developed based on the proposed approach. A user study is to be conducted to collect data that is required to allow the research team to test whether the design able to help the users to perform better in data analysis.

You are invited to participate in this project because you are one of the undergraduate or postgraduate students who have the financial background required in the user study.

PARTICIPATION

Your participation will involve a two-session user study at Y block level 6 or other agreed location. In the user study, you will use two different computer software to make stock investment decisions in two separate virtual stock markets.

You will participate in a 2-session user study. You will be using the prototype system in one session, and an alternative system in another session. Each session is approximately 120 minutes. In each of the sessions, you will go through a 30-minute guided training on the system (either the prototype or the alternative system). In the next 45 minutes, you will use the system to analyze a stock market and make an investment decision. Next, you are required to answer a questionnaire survey which approximately takes 10 minutes. Questions will include "to what extent do you think the software is useful in accomplishing the task", "to what extent do you think the software is easy to use", and "to what extent do you think you have understood the impacts of the interaction between factors and the stock prices.".

Your participation in this project is entirely voluntary. If you agree to participate you do not have to complete any question(s) you are uncomfortable answering. Your decision to participate or not participate will in no way impact upon your current or future relationship with QUT. If you do agree to participate you can withdraw from the project any time without comment or penalty. Any identifiable information already obtained from you will be destroyed.

EXPECTED BENEFITS

It is expected that this project will not directly benefit you. However, the project would benefit the information system research community by better understanding the supports that can help data analysts to make informed decisions based on large and complex datasets. Specifically, the findings from the project would provide useful understandings of how different design features will be able to effectively support the data analysts' problem-solving activities at various data analytics phases. These understandings would allow the research community and practitioners to design decision support systems that can better support users in decision-making process.

To recognize your contribution should you choose to participate the research team is offering a $20 voucher upon the completion of the user study. The participant who has the highest return of investment among all the participants will be offered a $300 voucher. Specific voucher can be requested.

RISKS

There are no risks beyond normal day-to-day living associated with your participation in this project.

PRIVACY AND CONFIDENTIALITY

All comments and responses will be treated confidentially unless required by law. Before releasing the thesis and any related publication, your personal identifiers will be removed from all data.

Any data collected as part of this project will be stored securely as per QUT's Management of research data policy.

CONSENT TO PARTICIPATE

The return of the completed questionnaire is accepted as an indication of your consent to participate in this project.

QUESTIONS / FURTHER INFORMATION ABOUT THE PROJECT

If you have any questions or require further information, please contact one of the research team members below.

Shiang Yen, Tan
Phone +61 3138 95007
Email shiangyen.tan@hdr.qut.edu.au

CONCERNS / COMPLAINTS REGARDING THE CONDUCT OF THE PROJECT

QUT is committed to research integrity and the ethical conduct of research projects. However, if you do have any concerns or complaints about the ethical conduct of the project you may contact the QUT Research Ethics Unit on [+61 7] 3138 5123 or email ethicscontact@qut.edu.au. The QUT Research Ethics Unit is not connected with the research project and can facilitate a resolution to your concern in an impartial manner.

Thank you for helping with this research project. Please keep this sheet for your information.

Section 1:

Questions in this section are related to your understanding of the stock market which you have just analysed. Please rate the following statements on a scale of 1 (very low) to 5 (very high).

	Very low	Low	Neutral	High	Very High
To what extent do you think:	1	2	3	4	5
Have you adequately identified the factors that are relevant to your stock investment task?					
Have you adequately identified the stocks that are potentially profitable from the stock market?					
Have you adequately understood the factors that are relevant to the stock market?					
Have you sufficiently understood the stocks based on the how the stocks qualify a combination of the relevant factors?					
Were you able to combine technical analyses into important knowledge about the stock market?					
Were you able to incorporate your judgements and assumptions into the understandings of the stock market?					
Have you have sufficiently understood the interactions between the factors in the stock market?					

Have you sufficiently understood the interactions between the factors in the stock market?					
Were you able to forecast the future price trend of the selected stock based on their current price trend?					
Where you able to forecast future price trend of the selected stock by considering the future movements of other factors in the market?					
Have you adequately evaluated the potential impacts of the stocks to be invested on your earning?					
Have you sufficiently assessed the potential impacts of the quantity of the stocks to be invested on your earning?					

Section 2:

Questions in this section are related to your opinion on your findings resulted from the analysis. Please rate the following statements on a scale of 1 (very low) to 5 (very high).

	Very low	Low	Neutral	High	Very High
To what extent do you think	1	2	3	4	5
The analytic results can be understood in the context of the task?					
The analytic results are supported by factual data and systematic techniques?					
The analytic results or its contents are new to you?					
The analytic results are not easy to be imitated by others?					
The analytic results are different from your expectations before the analysis?					
The analytic results are robust to uncertainties?					
The analytic results are derived based on realistic conditions and constraints?					
The analytic results are derived based on in-depth analysis processes?					
The analytic results are likely to help to solve the problem successfully?					
The analytic results are able to enrich your knowledge about the task?					
The analytic results are able help you in identify opportunities and avoid threats?					

The analytics results can directly provide inputs to the decision making?					

Section 3

Questions in this section are related to your opinion on the experience on using the software. Please rate the following statements on a scale of 1 (strongly disagree) to 5 (strongly agree).

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
	1	2	3	4	5
I am satisfied with the software.					
The software is useful.					
The software is easy to use.					
The software is easy to learn.					
I wish to continue to use the software in the future.					

Section 4:

Based on the stock market analysis activity that you just finished, choose the best response for the following statements.

	Very Low	Low	Rather Low	Neutral	Rather High	High	Very High
	1	2	3	4	5	6	7
The data covered in the task were very complex							
The task covered data that I perceived as very complex							
The task covered concepts that I perceived as very complex							
The task really enhanced my understanding of the data covered in the task							
The task really enhanced my understanding of the stock market in the task							
The task really enhanced my understanding of the stock market analysis							
The task really enhanced my understanding of the concepts in stock investment							
The system used during the task were very unclear							
The system was, in terms of learning, very ineffective							
The system was full of unclear design or interface							

Section 5:

Please select the best possible answer for each of the questions below:

Which of the following best describes how much you know about finance?

- I do not know anything about finance
- I know very minimal about finance *(I know how to read my bank statement and taxation statement)*
- I know some basic finance *(I know how car loan and house mortgage are calculated)*
- I know finance quite well *(I know how to prepare a basic profit & loss statement and balance sheet)*
- I know finance very well *(I know exactly what are present value, quick ratio, and working capital).*
- I know very much about finance (*I know how to calculate present value, quick ratio, and working capital and how are they related to each other).*

Which of the following best describes how much you know about stock investment?

- I do not know anything about stock investment
- I know very minimal about stock investment *(I know what and why is "buy low, sell high")*
- I know some basic stock investment. *(I know a few basic indicators for stock analysis)*
- I know about stock investment quite well *(I know how to apply and interpret indicators)*
- I know stock investment very well (*I know exactly what are the indicators measuring, and understand the fundamental logics underlying the indicators*)
- I know very much about stock investment *(I know exactly what combinations of indicators should I use for different scenarios, and clearly knows the reason for choosing such combinations)*

Which of the following best describes your actual experience in stock investment?

- I have never traded in stock market
- I have tried to trade for a few times
- I trade less than 10 transactions a year and still actively trading
- I trade more than 10 transactions a year and still actively trading
- I trade more than 25 transactions a year and still actively trading
- I trade professionally

Which of the following best describes the average duration how long you are willing / planning to hold your stocks?

- Daily basis (buy and sell on the same day)
- Monthly term (buy and sell within the same month)
- Short Medium term (hold between 3 to 6 months)
- Long Medium term (hold more than 6 months but less than a year)
- Long term (hold longer than 1 year but less than 5 years)
- Holder (longer than 5 years)

Which of the following best describes how you generally solve problems?

- Highly systematic and Highly intuitive
- Highly systematic and Lowly intuitive
- Lowly systematic and Highly intuitive
- Lowly systematic and Lowly intuitive

Which of the following best describes how much risk you willing to take?

- Very high risk – willing to risk losing more than 100% of the investment capital
- High risk – willing to risk losing 100% of the investment capital
- Medium risk – willing to risk losing 50% of the investment capital
- Low risk – willing to risk losing 25% of the investment capital
- Zero risk – willing to risk losing 0% of the investment capital
- Not sure

What is your highest qualification?

- Doctorate
- Master
- Bachelor
- Diploma
- Others

Which of the following best describes your occupation types?

- Professional (e.g. accountant, lawyer, doctor, scientist)
- Intermediate (e.g. manager, schoolteacher, engineer, electrician, farmer)
- Skilled (e.g. secretary, shop assistant, waiter, sales assistant, clerical worker, cook, carpenter, bus driver)
- Semi-skilled (e.g. agricultural worker, postman, telephone operator).
- Unskilled (e.g. kitchen hand, office cleaner, window cleaner)
- Retired
- Students

Which of the following best describes your area of expertise?

- Finance
- Economics
- Accounting
- Marketing
- Advertising
- International business
- Public Relations
- Human Resource Management
- Management
- Other _____

10.2 Appendix B – Ethic Clearance Approval

Dear Dr Taizan Chan and Mr Shiang Yen Tan

Project Title: Deriving actionable insights: A problem-solving approach to data analytics

Ethics Category: Human - Low Risk
Approval Number: 1600000553
Approved Until: 7/07/2017
(subject to receipt of satisfactory progress reports)

We are pleased to advise that your application has been reviewed and confirmed as meeting the requirements of the National Statement on Ethical Conduct in Human Research.

I can therefore confirm that your application is APPROVED.
If you require a formal approval certificate, please advise via reply email.

CONDITIONS OF APPROVAL
Please ensure you and all other team members read through and understand all UHREC conditions of approval prior to commencing any data collection:
> Standard: Please see attached or go to
http://www.orei.qut.edu.au/human/stdconditions.jsp
> Specific: None apply

Decisions related to low risk ethical review are subject to ratification at the next available UHREC meeting. You will only be contacted again in relation to this matter if UHREC raises any additional questions or concerns.

Whilst the data collection of your project has received QUT ethical clearance, the decision to commence and authority to commence may be dependent on factors beyond the remit of the QUT ethics review process. For example, your research may need ethics clearance from other organizations or permissions from other organizations to access staff. Therefore, the proposed data collection should not commence until you have satisfied these requirements.

Please don't hesitate to contact us if you have any queries.

We wish you all the best with your research.

Kind regards

Janette Lamb / Debbie Smith
on behalf of Chair UHREC
Office of Research Ethics & Integrity
Level 4 | 88 Musk Avenue | Kelvin Grove
+61 7 3138 5123 / 3138 4673
ethicscontact@qut.edu.au
http://www.orei.qut.edu.au